机器学习 VS 复杂系统

许铁 编著　　郭瑞东 审校

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书从跨学科视角来看待人工智能这个技术性的学科。围绕用数学模型预测未来这一主题，介绍算法，主要包括现在流行的机器学习和深度学习算法，以及算法要解决问题本身的复杂性。复杂的问题，需要复杂的算法，而算法设计背后的老师正是自然界的复杂性本身。最终，我们上升到自然界解决复杂性最有利的工具，即人类智能本身，让读者从神经科学的角度再次理解人工智能这个大主题，理解神经科学是如何启发人工智能的，而人工智能又如何帮助我们理解人类智能本身。

本书既适合具有高中以上数学知识的一般读者，作为他们了解人工智能和复杂系统领域的科普读物；也适合已经在人工智能领域工作的专业人士，使他们从工程视角之外的更大视角去看待这一领域，获得新的启发。

图书在版编目（CIP）数据

机器学习 vs 复杂系统 / 许铁编著. —北京：电子工业出版社，2018.8
ISBN 978-7-121-34410-7

Ⅰ. ①机… Ⅱ. ①许… Ⅲ. ①机器学习－普及读物 Ⅳ. ①TP181-49

中国版本图书馆 CIP 数据核字(2018)第 120900 号

责任编辑：安　娜
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　印张：12　字数：192 千字
版　　次：2018 年 8 月第 1 版
印　　次：2018 年 8 月第 1 次印刷
定　　价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

序

人工智能是大势所趋，这一主题也成为街头巷尾人们热议的焦点。本书作者的写作初衷是希望能在一个更大的视角下看待人工智能这个技术性的学科。虽然人工智能的技术更多用于工业界，但是其产生的根源，却与人类用数学模型探索世界和自己大脑本身的历程密不可分，因此，本书希望在这个大背景下，从高度跨学科的视角切入人工智能这个大主题。

人人都想预测未来，但是无论科技多么发达，预测未来依然是困难的，比如股市、自然灾害、一个月后的天气预测等，都很难做到精准。究其根本原因，在于无处不在的复杂性。复杂系统这门学科可以帮助我们理解复杂性产生的根源。

大数据时代，数据已成为人们最大幅度减少这种不可预测灾难的工具。然而过度信赖数据，往往会让我们陷入一种新的迷信，或者埋没在噪声里。

算法，主要包括现在流行的机器学习和深度学习算法，可以帮助我们最大程度地从数据里提取信息、剔除噪声，这也成为目前人工智能的基础。从另一个角度看，即从人类智能本身来看待这个问题，会让我们对机器学

习和深度学习的本质有着更新的理解，并发现两者内在的相似性，同时看到机器的算法和人脑的算法这两种相似性的根源，即世界本身的复杂性。

本书内容涉及机器学习基本方法、非线性动力学、复杂系统、随机过程、神经网络等正在深刻改变世界的学科，讲述人工智能和复杂系统的基本原理，以及它们是如何影响我们日常生活的方方面面的：小到微信里的语音识别，大到我们到底能不能预测经济危机或金融危机。

我们希望用化繁为简的笔墨，带领读者逐步理解这个大视角下的核心概念，掌握进入更多细分领域的语言和思维体系。我们的使命不是涵盖这一宏大领域的所有知识点，而是希望引发读者的兴趣，去寻找和了解这一激动人心的领域的更多进展。

本书既适合具有高中以上数学知识的一般读者，可作为他们了解人工智能和复杂系统领域的科普读物； 也适合利用业余时间充电，正在自学人工智能的工程师，作为他们搭建框架和进入专业领域的开胃菜； 还适合已经在人工智能领域从业的专业人士，使他们从工程视角以外的更大视角去看待这一领域，获得新的启发。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/34410

目　录

第一部分　复杂性

第二部分　机器学习

第三部分　神经网络

第四部分　宇宙间最复杂的就是我们的大脑

第五部分　人工智能应用谈

第一部分　复杂性

1 复杂系统

何为复杂（阅读难度★）

复杂系统用来阐述模式的产生。现实生活中模式无处不在，例如，同类型的商店往往比邻而居；分久必合，合久必分，热恋过后不是分手就是回归平淡。甚至那些我们不认为是模式的，比如生命过程本身，其实都可以看作自然界中模式涌现的过程。这些模式往往可以归结为组成系统的个体通过简单相互作用达到某种非同寻常的集体的现象，复杂系统用动力学的思想阐述了这一过程的发生。

（1）系统可以被分解为比自身简单许多的个体。市场→人，生物→细胞。但是个体和系统之间却呈现了完全不同的属性，用完全不同的特征描述。比如，人具有的生物属性是吃喝拉撒，而由很多人组成的社会系统，就具有了国家、政府、交易市场等这些特征。

（2）个体之间的相互作用反过来受到系统制约。比如，人与人交换物品，需要依靠共同的媒介——钱，而依据的标准是大量交易行为中涌现的价格。这些制约因素往往是在个体的相互作用中自发涌现的。这些制约可以看作一种宏观秩序或者组织，使得相隔遥远的个体通过它联系起来。

一些非常简单的个体之间的相互作用规则就可以诞生出非常复杂的宏观范式，这一现象最早受到物理学的启发。一个里程碑式的模型被称为 ising 模型，这个模型利用不同条件下简单晶体的演变阐述了最复杂的物理问题——相变，在后面会继续介绍。

（3）系统与外界环境的作用可以用信息的流入和流出描述。复杂系统在外界信息发生变化时的反应是描述复杂系统属性的根本，很多能够稳定存在的复杂系统都具有相对外界环境自发调整自己结构的能力。

最典型的复杂系统的例子就是股市，大家都想预测股票价格，但是很难预测，这是因为股票价格本身即是大量交易者买卖之间涌现的一种宏观秩序，它与公司的业绩并不明显相关，公司业绩对股市的影响更像是进入交易市场这个复杂系统的一个外部信息，外部信息影响交易者的心理但不决定他们的行为，最终股票的价格是由交易者间的相互作用（博弈）决定的。如果要研究股市，更多的是研究如何根据交易者所透露的蛛丝马迹推测其可能的行为。即使你能够跟踪每一位交易者，也由于上述几个原因，股市依然很难预测。

金融市场的不确定性的根本来自于复杂系统，这使得我们常常低估风险。比如次贷危机，我们依据的假设很多时候是把市场看作独立作用的部分，每个部分的风险是独立的。事实上市场的每一部分并没有独立，正是它们之间的互相关联导致次贷危机的。

为何复杂（阅读难度★★）

复杂系统的复杂，是由哪些因素影响造成的呢？

作用（关联）

不是单体的特性，而是单体是如何相互关联形成组织的。因为这类系统共同的特点是长程关联。关联往往导致 1+1>2 或 1+1< 2，或称为非线性。市场中出现的价格是受网络相互作用导致的，我们都受到邻居的影响（Herding Effect），相互作用非常重要。比如，神经元就是通过相互作用构成神经网络来处理信号的。

相互作用使得物理系统无法轻松地由整体拆成部分，非线性即其根源，那么还有

什么方法可以解决这个问题？

一旦这种相互作用的维度增加，就会产生一个新的现象，即混沌。混沌说的是由于系统内自由维度的增加，系统的动力学属性不再归于闭合轨道，而是开放或成为不可预测的轨迹，初始条件的轻微变化在未来的影响远未可知，想想真是脊背发凉。

相互作用导致协同效应。两个人在一起可以是 1+1>2，也可以是 1+1<2，但基本不会是 1+1=2，前两者可以看作非线性的体现。比如为什么会有公司，那一定是某种合作导致 1+1 大于 2 的效应使得公司可以产生。

反馈

复杂系统多用于描述一个系统的时间变化过程，比如市场价格的波动、神经网络随时间的活动等。研究这个时间变化过程，往往要考虑此刻的结果对下一刻系统输出的影响。股市的反身性就是反馈的一种。

反馈分为正反馈和负反馈两种，负反馈导致定点平衡态，正反馈导致不稳定性，如雪崩、股市崩盘等。在所有复杂系统中，都有正反馈和负反馈两种状态。反馈带有回路的概念。一个单元通过相互作用传递给另一个单元，反过来另一个单元又可以把信息传递回来。反馈往往是指此刻活动对下一刻活动的影响，比如市场价格。市场价格永远围绕均衡波动，价格高，导致市场买的人变少，买的人少后又导致价格降低，这是典型的负反馈。负反馈把系统维持在稳定位置。$dx=-x$，这是负反馈。

相变

这是复杂系统的第三个重要特质，而且是组织形成的核心。当系统主导反馈的性质发生变化时，则经历一个相变。

相变在自然界和社会中无处不在，自然界中的相变既包括冰和水之间的转化，也包括磁铁从一种相到另一种相的变化。物理中相变的典型例子是磁铁的 ising 模型。

磁铁这个东西，并非总具有磁性。那么具有磁性和不具有磁性的磁铁有什么区别呢？

磁铁有两个相，一个是组织成分均匀一致（有序）的状态，另一个是无序而混乱的状态。虽然它们都是由铁原子构成的，但铁原子只有在有序排序时才会产生磁性。

当需要对外发挥一种作用时，就需要齐心合力，而无序的铁原子使得每个磁针的磁性相互抵销了。

这里就建立了相的概念。而相变，就是通过外部变量使得整个系统从一个相到达另一个相的过程。相变理论是复杂系统研究的重要对象，我们知道，磁铁有的有极性，有的没有极性。研究磁铁极性变化的模型被称为 ising 模型，说的是 paramagnetic（无磁性）到 ferromagnetic（有磁性）的变化。这里影响一个系统相变的主要因素有两个：一个是熵（无序性，系统信息的缺失），另一个是某种趋同的效应。

在铁磁物质里，每一个原子都有极性，平行排列的原子具有指向相同方向的趋势，而熵无序时则破坏了这种效应，两种力量互相争夺。在较高温度下，熵的作用占主导；而较低温度下，有序的趋同的力量占主导。在某个温度下，磁体的原子从无序的状态过渡到完全有序的状态。在完全有序的状态下，整个磁体就显现出对外的磁性。

这里，我们可以控制的外部变量就是温度。温度越高，熵越大。当温度为 0 时，系统自由能最小的状态是一致有序的态，温度升高，无序的态的自由能逐渐减少，直到某个点，成为更有优势的态。类似的还有水到冰的相变，也是在某个温度上，无序和有序交替，这称为临界。所谓临界，就是相变时候的状态，因为这个时候最为特别。临界点上的系统属性特别复杂，统计上我们经常会看到肥尾分布或类似肥尾的分布，这样的分布无处不在，比如股市价格波动、工资分布（帕累托）等。

临界极为重要，为什么？

因为系统在临界点上的属性特别复杂、丰富且有趣，而且，更重要的是，大部分和我们息息相关的系统事实上都在某种程度上处于临界态（或靠近临界态），包括大部分的生物系统和经济系统。

更多关于临界态的内容，可以参考《大自然如何工作》这本书，知乎上的大 V 傅渥成的回答。

> “此外，所谓涌现，是在刚才讲到的作用、反馈以及自组织临界基础上得到的，系统从微观到宏观，性质而非数量上‘从零到一’式的改变。”

最简单的例子是路，所谓人走得多了就成了路，森林中交错的小径是大量人走过所涌现出的一种现象。涌现性和相变点也有千丝万缕的联系，有兴趣的读者可以关注自组织临界（Self Organised Criticality）的理论，去查看更多这个领域的知识。

复杂系统的元素很多，而且元素之间均有相互作用，最好的刻画方法就是复杂网络。

2 用复杂网络看世界经济

（阅读难度★）

图 2.1 描述的是一个叫作产业森林的念。产业森林描述了一个由种类繁多的农业和工业产品组成的关系网。每一个产业即网络里面的节点，就像产业森林里的一棵树，例如图上标出的谷物、采矿业、电子产品等。这里说的是，如果把这些产业描述成复杂网络的一个点，那么我们会对这件事有一个全新的理解。

如何把产业和复杂网络联系起来呢？首先，要做出一个网络结构，我们必须有距离和连接的概念。事实上，不同产业之间不是孤立的，这里所说的产业之间的相似度，其实就是构成它们的技术或生产资料要素的相似性。我们可以规定超过一定相似度阈值的两个产业就可以被联系起来，这样，即可得到不同节点两两相连的矩阵，画出来就能得到产业的复杂网络图。

森林中心是高级工业品，森林边缘是农产品和原材料等。森林中心树木密集，而森林边缘树木稀疏。值得注意的是，树木密集的地方树木间也更盘根错节（连接密集），而树木稀疏的地方连接也越来越少。在森林的边缘，有些树木基本就是孤立的，比如说奶牛生产。这个图颇像城市的交通图，市中心路网交错，而到了郊区的小村，基本就到了路的尽头。

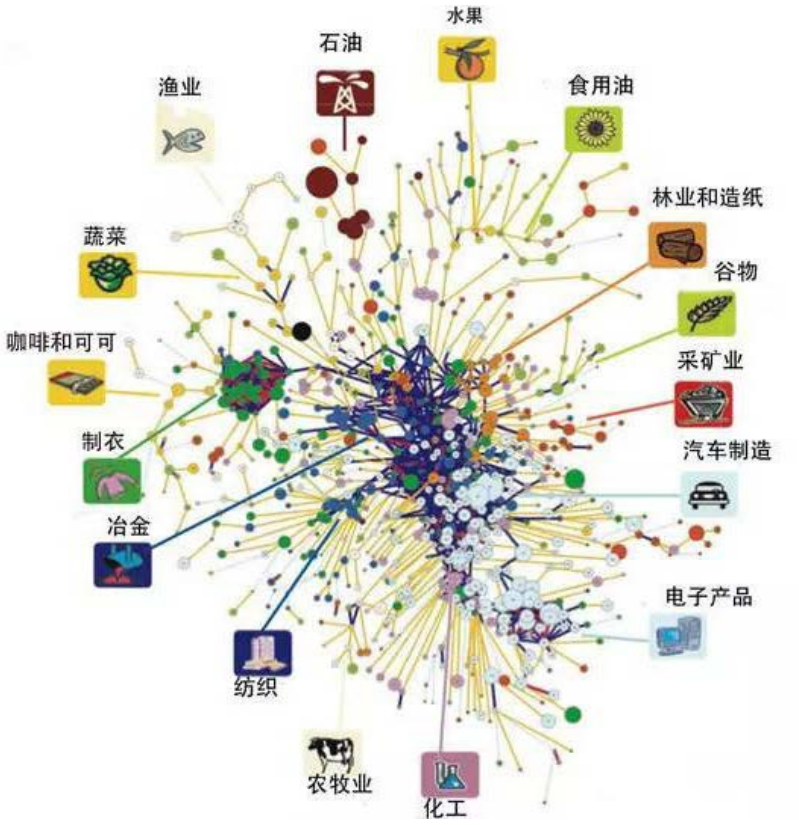


图 2.1

这个网络如何理解？

第一，每个产业间的连接刻画出了两种产品间在生产过程中的联系，外围的产品是内侧产品的原材料，当然内侧产品也可以给外围产品提供服务。第二，连线的颜色对应两种产品的相似度。何为相似度？这个概念可用来刻画生产它们时需要的生产要素的相似性（如劳动、技术、资本……）。例如，梨和苹果的相似度就很高，如果一个地方适合栽种梨树，那么它也往往适合栽种苹果。因为梨和苹果的生长要素差不多，同样的气候和土壤，都需要大量劳动力投入。节点的颜色被用以描述产业的大类，比如资本密集型都用蓝色（图 2.2 中的⑦）表示，劳动密集型都用蓝色（图 2.2 中的⑧）表示。节点的大小代表年交易量，如图 2.2 所示。

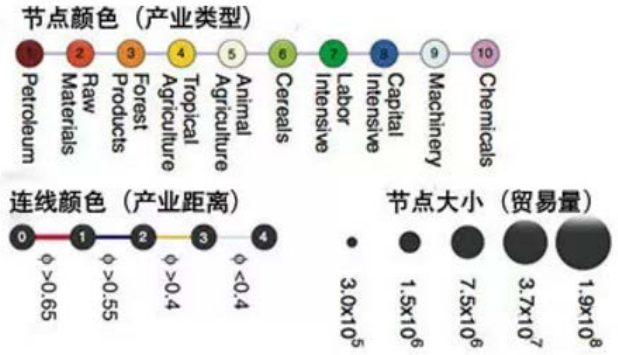


图 2.2

重新回去看产业森林的图，就会立刻明白，森林的边缘树木稀疏而中心树木繁密（树木的间距即产品的相似度，相似度小则间距大，注意树木的间距是用颜色表明的，原图见下载资源）。边缘的树木果实稀少（小圆，贸易量小），而中心的树木硕果累累（大圆，贸易量大），石油产业除外。

换句话说，这张图对全球的所有产业都做了一个聚类，越靠近中心的产业附加值越高，边缘的则与之相反。

这样一个网络结构究竟有着何种重要意义呢？传统理论认为这样的一个结构图没有什么用，因为条条大道通罗马，只要经济在积累，总有机会到达果实丰硕的森林中心。

而复杂网络理论则告诉我们一个完全不同的故事。首先，一个企业往往用它所占据的较为优势的产业来表示。一个企业的发展，被称作从森林的某些位置向其他位置跃迁的过程。可以把企业比喻成产业森林里跳越的猴子，它可以从一棵树跳向另一棵树（产业升级）。当然猴子的目标都是朝着果实丰硕的中心去的。以往的文章往往重视研究猴子，如果一个猴子长期无法到达中心，一般认为是猴子的问题。一个预先的假设是猴子只要跳得足够好，总能从一棵树跳到另一棵树并到达森林的中心。

下面我们从森林的结构来阐述为什么有的猴子能够到达森林的中心而其他猴子不行，因为森林的结构是很重要的制约因素。

网络的性质至关重要，因为有些网络结构，任何一个猴子经过足够长的时间都可以到达森林的中心；而对于另一些网络，猴子能否到达中心，则完全取决于它初始所在的区域，在某些区域，即便它有三头六臂也无法到达森林的中心。

前面提到过，森林的边缘树木如同独立的孤岛，而中心的树木密集（产品相似度大），如同大陆。猴子的跳跃能力是有限的，猴子从森林外侧向中心跳跃不是一件容易的事情。当猴子所在树的周围树木过少且树木间距过大时，猴子就无法跳到下一棵树上。森林中心的猴子可以很轻松地在树木之间跃迁并摘取丰美的果实，而边缘的猴子则没什么选择。

另外一个发人深省的现象是猴子能否进入森林的中心和它所在的初始位置以及它选择的跳跃方向十分相关。如果它开始所在的位置恰好树木间距不太大，而且存在能够到达森林中心的道路，那么它将很有机会进入中心，反之则很难。即便如此，如果一只猴子没有选择好方向，不小心跳到了独木上，四周没有其他树，那么它被困在那个地方的概率就很大。这里其实也有不少例子，当然只考虑森林的结构肯定是偏颇的，但森林的结构的确是很多例子背后的隐藏变量。

图 2.3 和图 2.4 是产业森林的局部截图。从图 2.3 可以看到，劳动密集型工业（汽车制造）和电子产品工业是一条由森林边缘通往中心的捷径，品种繁多，相距很近。从图 2.4 中我们可以看到热带作物（咖啡和可可）一直处在森林的边缘，与森林中心差距较远（仅橡胶与工业品联系较密切），各种热带作物间也距离较远。

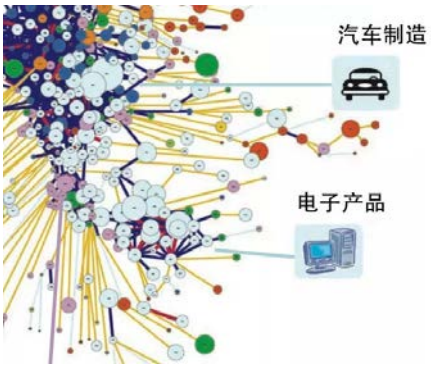


图 2.3

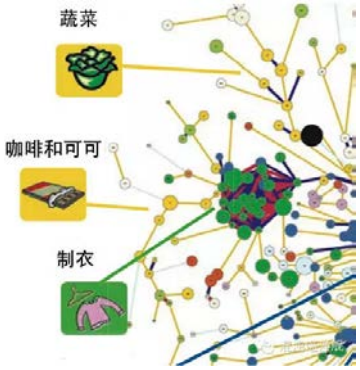


图 2.4

3

风险管理策略之复杂科学视角

黑天鹅效应

曾经一段时间，主宰统计世界的是一个叫作“高斯分布”的函数，它的英文“normal”含有正态、标准之意，说的是决定事物整体性质的是它的平均，比如说你可以用 1 米 7 代表整个中国人的身高。基于高斯分布的各类数学工具主宰着庞大的金融帝国上百年，却成为 21 世纪初金融危机的罪魁祸首。这背后的缘由，正是复杂性主导的黑天鹅效应。

在黑天鹅出现之前，天鹅湖里的天鹅都是雪白，你可以想象那种天蓝色的湖面上飞起千万只白天鹅的感觉，远远看去如同乞力马扎罗的雪。于是我们把白当成天鹅的标志，以趋于 100%的概率预测天鹅皆白。直到有一天湖面上飞过一只纯黑的天鹅，宛如来自世外，但它却打碎了我们的白天鹅之梦。从而我们明白，在生物世界里特例才是本质，而不是平均。特例总会以比你预想的还要大的概率出现，而把之前的理论打得粉碎。

经典物理的世界是平均数的世界，细节和特例都可以滤掉。但一旦进入生物主导

的领域，它们就变成了王道。

在进入混乱的生物世界之前，我们先来看一下高斯定律主宰的“白天鹅之舞”。

高斯定律与大数定理——平均的力量

我们经常用平均数表达事物的总体状况，对于做统计的人，平均数几乎成为信仰，我们往往已经忘记了这种信仰背后的基本假设——高斯分布，只有在我们统计的事物呈高斯分布时，平均数才能够代表事物的属性。

预备知识：加和等于平均。平均数的运算依赖于把很多的数据相加后除以数据的个数。样本在平均数附近的偏差大小用标准差表示。经典理论告诉我们，样本的容量越大，平均数就越能代表所研究的群体，图 3.1 所示的高斯吊钟曲线，中间的竖线指代平均数，底下的小横杠标注的是标准差。

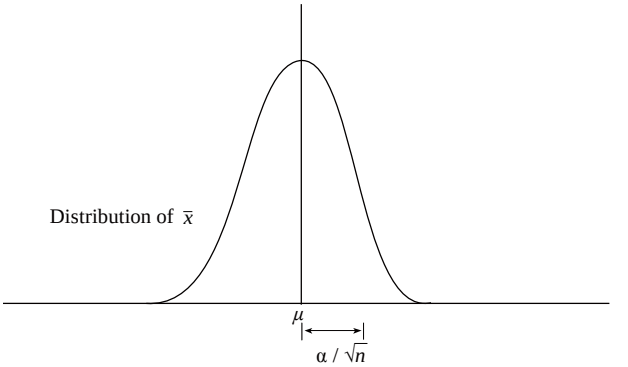


图 3.1

高斯定律让我们看到了加法的威力。对于一个随机事件，比如掷骰子，虽然每一次取得的结果从 1 到 6 完全无法预测，但是如果掷一万次，把每次掷的点数加起来就能得到一个可以被越来越精确预测的数。这个结果可以被一条高斯曲线描述，它具有两个特征量：平均数和标准差。平均数描述总体趋势，而标准差则告诉你不确定性的大小。随着加数的增多，标准差在平均数面前越来越微不足道，直到可以忽略不计，或者说通过无穷加和，一个随机事件成为确定事件，我们越来越精确地得到平均数。这条法则叫作大数定理（Law of Large Number)，如图 3.2 所示。

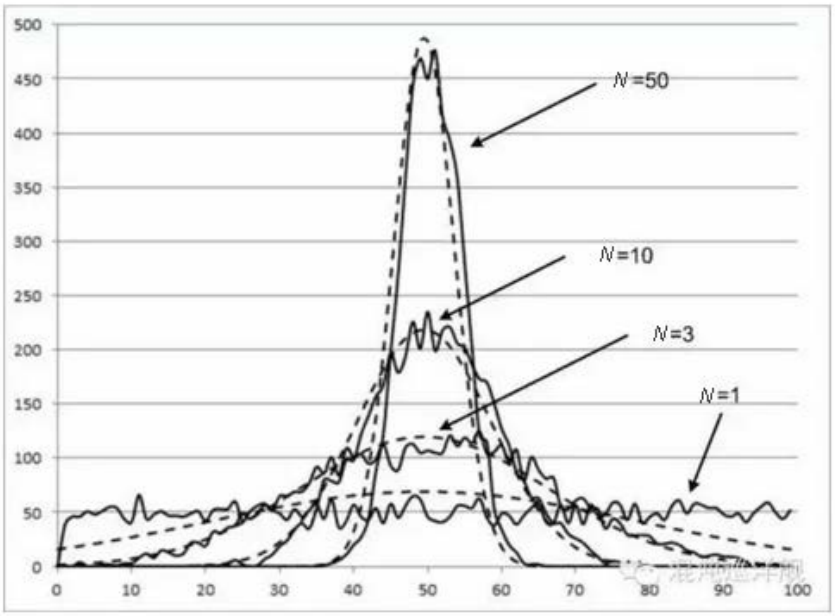


图 3.2

我们看到随着样本总数 N 的增加，钟形曲线越来越瘦（对平均数的偏离减小），想象一下，当 N 无限大时，我们就会得到一条竖线：代表我们以任意精度都能得到平均数，不确定性被消除。

大数定理的威力，在于它使得一个确定性的世界可以在庞大的不确定性之上产生。就好比明天太阳升起、春天花儿会开这种事，我们知道不发生的概率几乎为 0，其实都源于高斯分布和大数定理。因为太阳升、花儿开是组成太阳和花儿无数的原子和分子共同作用的结果，一旦沾上“无数”“共同作用”，大数定理就可以精确地保证事物一定会发生。

正态分布和大数定理是所有确定性的根源，因为我们的可见世界就是无数不确定的微观因素不断加和的结果。

对于细节重要性的启示：它告诉我们当决定事件的因素足够多、试验的次数足够大时，微小的细节不再重要，因为它们在巨量的加和中被平均掉了。

不过不要高兴得太早。

高斯定律背后的陷阱

A. 细节因素要独立

看上去有点抽象，其实说的是那些加数组成事物的要素不能私下暗自沟通，例如，如果你认识的所有女性私底下串通起来说你很好或很坏，那么你约会的人的数量不会超过理想平均数，因为所有的女人其实都取得了和你最开始约会对象一样的想法，你得到的只是放大的标准差，你第一次约会里的随机性被放大成为一生的结果。初始条件的影响被放大，正如亚马逊森林的蝴蝶扇扇翅膀，引起大西洋上的一场风暴。

B. 时间平移不变形

举个简单的例子，如果你投掷的时候骰子被人换掉了，变成一个加了机关的骰子，每一面都是一点，而且后面又经常被时不时地换掉，那么你永远得不到稳定的平均数。如果你仍是按照高斯定律做加法，指望最终会赢得平均数给定的钱数，那么被骗的概率就很大。

大数定理是我们认识随机世界的基础，它告诉我们确定性如何从偶然性的基础上浮现。但是它就如同牛顿第一定律和理想气体模型，光滑水平面和无相互作用的基本粒子在真实的生物世界如同幻影般不存在，虽然我们的确在某些时候会得到一些趋近的情况。

黑天鹅效应与幂律分布

高斯定律和大数定理保驾着庄严的理论物理世界，在这里，好好学习就能天天向上，灰姑娘一定会遇到王子。但是，黑天鹅还是摧毁了童话。

黑天鹅的本质是个体对总体、细节对全局产生决定性影响。当湖面出现一只黑天鹅时，整个天鹅群体的属性就发生了变化，一个纯白的世界霎时变成中灰。当然，这里更多的是看到特例的影响。

用高斯的正态的观点来看，黑天鹅出现的概率本来可以忽略，因为我们之前已经统计了巨大的白天鹅样本，但黑天鹅还是出现了，是我们的运气特别不好吗？错。但错的不是你，而是正态分布。

在生物世界里，主导的是幂律分布（Power Law），其实它也正如 power 的英文

原意，与权利和财富有关（见图 3.3，帕累托分布局部决定整体的象征，Mandelbrot Set 分形结构的代表）。幂律分布的数学表达式简洁无比，不同的幂律分布只体现在幂指数的不同上。它与高斯定律的本质不同在于，高斯正态分布下那些概率小到可忽略的事件，在幂律分布下却没有那么罕见。在幂律分布的观点下，黑天鹅的出现是可以理解的。罕见的黑天鹅不仅来到，而且决定着全局。

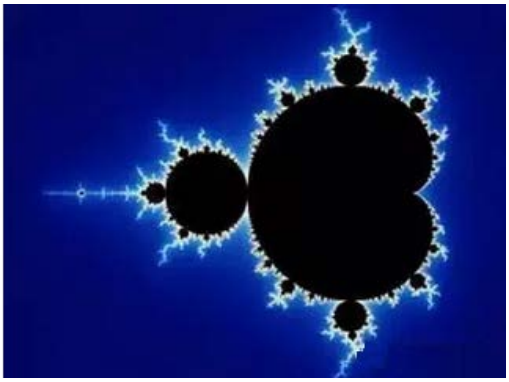


图 3.3

局部的特征与全局特征具有自相似性，幂律正是它的数学表达。此处由于篇幅所限，不再继续，感兴趣的读者可自行查找资料。

高斯分布与幂律分布的对比如图 3.4 所示。幂律分布最显著的特征是它的长尾，表示那些在高斯分布下的微小概率事件并非那样罕见。

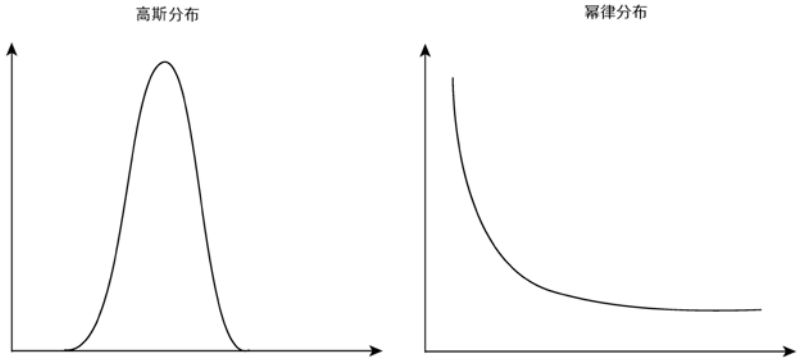


图 3.4

严重偏离平均值的事件在幂律分布下不再是偶然，并且掌控全局。例如，帕累托指出的社会财富的二八定律，20%的富人掌握 80%的财富，这个少数赢家通吃（Winner Take All）的规律几乎统治着市场经济下的各个领域。还有生态系统里面的大鱼吃小鱼，鱼的尺寸也是幂律分布。它们均体现了在这些体系内元素间存在的全局关联，你我互吃构成的因果链。幂律分布在经济学中的显现就是帕累托分布。

为什么黑天鹅影响如此之大?现代物理中的相变理论给出了震撼有力的答案。下面用一个具体的例子点明——雪崩。雪崩是山顶大面积的雪体坍塌，本来要推倒一座雪山几乎是不可能的事情，雪崩符合经典的黑天鹅事件的定义，按常理几乎不会发生，一旦发生即致命。为什么在现实中我们会经常听到雪崩的事故呢？因为雪崩的诱因与它的影响相反，非常的微小，可能是一粒小石子打到雪山上，或者一个人在喊话，这些诱因没那么罕见。这些微小因素在绝大多数情况下都对雪坡毫无影响，但是在一种情况下，即雪体的临界状态，就会发生雪崩。

临界状态是一种脆弱的平衡状态，维持雪体凝聚在一起的力量和使雪体瓦解的力量几乎相等，但是只要天秤稍微倾斜便万劫不复。你在庞大的雪坡上投一粒微小的石子，石子的作用力不是被局部的雪体吸收，而是扩散到整个雪体，如同压死骆驼的最后一根稻草，使平衡整体倒戈。

临界状态使得黑天鹅成为决定性的力量。

雪崩理论的核心是临界状态下细节的作用被无限放大（正反馈）。一个本来只限于局部的小因素在临界状态下扩散到全身。雪崩的理论遍布各个领域，例如地震、股市崩盘、金融危机。

在一场势均力敌的战斗中，任何个体微小的作用都可以被放大并左右战局。比如在一场两边力量均等的拔河比赛里，某支队伍的一个成员手机响了，他一慌神松了口气，而这一效应又导致慌张情绪在全队蔓延，结果由于一个手机铃响而输掉了一场比赛。所谓丢失一个钉子，坏了一只蹄铁；坏了一只蹄铁，折了一匹战马；折了一匹战马，伤了一位骑士； 伤了一位骑士，输了一场战斗； 输了一场战斗，亡了一个帝国。越是势均力敌的高手比赛，越要拼运气。

生命洪流的本质是一种特殊的相变，因此与生物有关的事情，包括生物的历史和

我们人类的历史以及我们每个人的一生，都发生在临界状态，那个小小的雪崩的状态，那个不可预见的细节决定了全局的状态。既然明天还活着，黑天鹅就会起飞。这对我们日常应对风险的策略，具有深刻的启示。

主动应对风险（阅读难度★）

世界如此多变，我将何以应对？

庞大的计划有用吗？No，试错万岁

如果让你回到过去重新玩一遍人生，一般人都会提出我可以怎么样，那就无敌了。如果按你想的那样来是否更好，我不确定。但可以确定的是，一定没有多少人觉得自己的生命是按照一条最优轨迹进行的。我们的生命过程就像一个盲人摸象的过程，无论站在哪个时间点，你的信息量都非常有限，根据非常局部的信息做出最优选择的机会几乎为 0。庞大的世界，复杂的历史，我们都捆绑在自己的路径上，在黑暗里瞎摸。

面对这种情况，最好的办法就是不停试错，任何过度思考和过度计划都是多余的。

本来就是一片黑暗，多想只会耗费心神。你应该通过快速摸索，增加你对周边信息的把握。每一次错误，你都可以根据它矫正你对世界的判断，这样，几轮之后，你得到正确选择的概率将会大大增加。

复杂系统的随机性和非线性，导致我们对它的预测无能为力，物理里称之为混沌。三个物体的相互作用就可以出现混沌，更不要说由无穷多非线性元件组成的复杂系统了。南美洲的蝴蝶扇扇翅膀，就可以引起北美的一次风暴。

一个利用试错进步的典型例子是市场经济。

市场经济把经济活动的自主权还给个体，虽然每个个体都不是很聪明，但是它们都有一个特点，知错就改，唯利是图。它们所主宰的经济，试错和纠错能力都是超强的。其结果是，短时间内资源分配就接近了最优化，虽然还有点波动，但仍超过世界上最厉害的经济学家的预测能力。

市场经济是反脆弱的，每一次意外的小概率事件，即黑天鹅事件发生时，它都可以调整过来，并且变得更加成熟，自由市场调节平衡的能力十分惊人。

另一个利用试错进步的典型例子就是生物系统。

生物系统可以算是已知的最复杂的系统，大自然在创造它时，并不需要一个精密的计划表，而是用一个笨办法，即选择算法。第一，生物系统可以通过变异无限试错；第二，只有能够适应外界变化的个体才可以把它的基因传给后代。两个简单的办法，加上一定的运算时间（进化史），就产生了无比有效却能够抗击各种自然灾难的生物体。

无论是经济系统还是生物系统，都是系统内的自由单元通过不断试错进步的，自由单元本身不知道下一刻的命运，而系统整体却坚定地迈向光明。如果束缚了系统内个体的自由，则试错的威力无以发挥。因此管理这些系统时，唯一该做的就是维护这种自由。

大自然的算法，你也能够掌握，生活中的进化算法就是多给自己些选项，多准备些备胎，然后快速试错，积累经验，稍作改进，继续试错，继续改进……

以无常对无常，要用不同幅度调整自己

复杂性告诉我们，世界是高速变化的，而且这种变化往往无序且不可预测。如果你想跟着世界变化，那么你永远是落后的。因为当变化的趋势路人皆知的时候，也是趋势将要变化，新一轮无常开始要肆虐的时候（看看股市）。因此，个体在面对环境变化的随机性时，要主动寻求变化，以动制动，在运动中搜索信息，调整步幅。

自然界应对无常环境进行的一种典型运动方式是先确定一个大的区域，然后做小范围的改变（试错），如果得到的反馈信息是不利的，就快速做出一系列大幅度的调整，直至达到一个比较有利的位置，这样的变动周而复始。这是一种应对无常最佳的适应方法。如果环境丰饶，则不失良机。反之，又不至于因过分执着于不够好的机会而被困死。

一个典型的例子是鲨鱼觅食。鲨鱼在鱼类丰富的环境进行小步伐的随机游走，只要不停地游动就可以吃到最多的鱼。但当鱼类相对不足的时候，鲨鱼就会进行大步伐的跃迁，这种跃迁也是随机的，却具备一次改变较大的特点。小步伐的随机游走比较容易穷尽开采一个地方的资源，但不容易到达较远的地点，而大步伐的跃迁却有利于开发新的领地，寻找新的食物来源。

自然环境对生物来说就是食物分布，最大的特点就是不均匀，要么大量来（自助餐），要么什么都没有（大饥荒），而这种变动往往不可预测，这就是鲨鱼选择这种觅食策略的原因。即小范围调整试错，若是找不到食物，就走一大步。

总结：自然界的变化幅度灵活多变，人也可以具备这种特点，先给自己制定一个大方向迈出去，在一段时间稳住这个方向后再进行小范围的自由调整，当发现趋势变化的苗头或者觉得自己走错了时，就大步伐改变。

对特别犹豫不决的人还有一个不错的故事。

在《反脆弱》这本书中有这样一个例子：一头驴子又饿又渴，前面是一条河，后面是草原，但是都要跋涉一公里。它犹豫不决，就好像二力平衡下的物体，死路一条。这个时候，唯一能够解救它的是随机走一步。结果朝哪里近，就去哪里。这就是随机运动的美妙，整个宇宙都可以理解为在随机运动下导致对称破缺（有序产生）的过程。

用否定法抓大放小，否定最致命的，放开其他的

在无常的世界里，每一个未来事件都最好看作一个分布函数，而我们一定要注意的是那些最致命的结果，如果最致命的不出现，活着就有机会等到胜利！

一些人总是拼命地注意小节，桌上任何东西都要摆放好，却忘记了关煤气灶。注意每个小节会消耗大量能量，过于注意小节反而可能疏漏那些致命的威胁。因此我们应该按照风险的轻重缓急来相应地分配注意力。

事实上，对于复杂系统，我们只需管理某些对系统产生核心影响的事件，而对其他事件放任，让自然来管理。

抓大放小，就是在能量有限的情况下，专注于做重要的事情。若是把精力过多地放在微小因素上，就会无暇顾及核心因素。而微小因素往往会在恰当的时候自发解决（比如一些不重要考试前夜的疯狂补习）。老子所说的无为而至，就是指大自然早已给我们设计好了节能优化模型，把一些事情交给自然，剩下的事情才可以尽力到底。

把握复杂系统，让它朝着人为意志方向往往适得其反，但是否定一些最坏的东西却是切实可行的。例如，在选择配偶上，每个人都应该问自己，最不能接受对方做什么？如果这恰恰是对方无法避免会做的，就可以确定不该在一起，至于其他的小的坏

习惯则可以置之不理。

杠铃策略——风险对冲

前面谈的策略都是比较消极被动的，好了，现在是我们主动出击，反向利用“无常”和分布函数进行获利的时候了。

什么是杠铃：两头重，中间轻，其实它就是无常的化身——幂律函数的缩影。幂律函数有两个特点：一个是“大头”，另一个是“长尾”。大头是较高频率，但影响微小的事件，而长尾则是较低频率，但对系统产生重大影响的黑天鹅（积分发散），在长尾理论中被经常提到的幂律函数如图 3.5 所示。

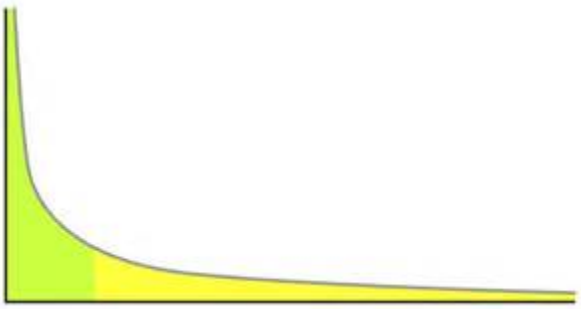


图 3.5

杠铃策略，就是同时把握大头和长尾，利用分布函数获利。这和中庸法则恰恰相反，中庸法则是远离极端，而对冲则是利用两个极端的综合博弈实现平衡。

下面利用复杂系统的非线性动力学分别介绍弱杠铃策略，以及仅仅利用分布函数形状的策略和强杠铃策略（杠杆原理）。

弱杠铃策略：风险组合

最简单的应用莫过于高风险和低风险事件的组合。比如，一个人应先有一个稳定的职业，然后再去做一些高风险的投资。

人是无法承担无限风险的，一个领域的风险，总要由另一个领域来吸收。如果你找一个风情万种的法国美女做老婆，那么你很可能就要承担被戴绿帽子的风险。相反，如果找了吕后那样稳定持家的老婆，那么你就可以在事业上挥霍或冒险，因为总有个

安稳的后方。

比如要做一件有压力的大事，那么你就需要一些特别简单的爱好让你能够在工作之余吸收进去，那个大事和无聊的爱好恰好是哑铃的两端。另一个典型的例子是如果生活特别枯燥，比如常和数字打交道，那么就需要做一些疯狂的事情释放，看看《华尔街之狼》就明白了。

总之，当你人生从事的事业中包含了具有相反极端属性的事情时，那么你就可以在风向逆转的时候通过对杠铃另一端的拥抱得到解救。

杠铃策略的另一个典型应用是人的知识结构，最有效的知识结构亦呈现幂律分布，专一的技术是头，广博的知识是尾。专一技术是人立足江湖的必杀技，但是在很多特殊的关键时刻，广博的知识又起决定作用。在巨大的未知性面前，仅有专一的技术往往是脆弱的，就像溺水的电脑专家，决定他命运的不是电脑技能，而是是否会游泳。

幂律之所以广泛存在，也在于其自身结构具有符合造物法则的反脆弱性。好比历史总是由少数杰出人物引领的，而多数人民提供基础，两个条件缺一不可。

强杠铃策略之杠杆原理：风险对冲

笔者从金融上的对冲基金得到启发，总结了一个较强版本的杠铃策略——对冲法。前面的风险组合强调互补，而风险对冲则强调相反相成。简而言之，就是一种事物的风险，恰恰构成另一种事物的机遇。杠铃一端的损失就是另一端的收益，当一端向下时，另一端恰好向上，好比跷跷板，也好比杠杆。聪明人利用这个杠杆，把生活中向下的波动转成向上波动的契机。

对冲的基础其实是事物的非线性，如果你同时买入分布两端的事物，而这两个事物又存在反向关联（当 A 下降时，B 有上升趋势，或反之），最关键的是，这种关联是非线性的（A 的下降不等于 B 的上升，下降总是小于上升），一端小的下降总会引起另一端较大的提升。

如果哑铃两端的事物具有凸函数的关联性，即 A 的减弱引起 B 的上升，而 B 的减弱又引起 A 的上升，并且两种变化的和呈上升趋势，那么同时买入这两种事物，

我们就实现了风险对冲，或者说反脆弱性。因为在任何情况下你的获利都为正。

让我们看看对冲基金是怎么操作的，对冲者同时买入一个行业内较优和较略的几只股票的卖空期货（当股票的跌幅大于预期时，就收益），这就是笔者所说的占据分布的两端。当行情见长时，较优股票的收益将大于较略股票卖空的赔损；而当行情见跌时，略等股的卖空收益将高于优等股下跌的损失，从而实现风险对冲。

如果你发现一对具有反向相关性的事物，而且这种相关性具有凹函数性质，那么你就可以做一笔好买卖，因为凹函数有在波动中受益的本性。

例如，假设李白成了宰相，那么很可能少了一位文学泰斗，而多了一个我们不知道的唐朝宰相；失眠的时候就博览群书，那么很可能就成为了一个有智慧的人；事业不顺找一大堆朋友喝酒，然后叫了一大堆好友一起创业很可能就成功了。

笔者做一个大胆的推测，这个世界能被记住的所有成功者都是通过某种杠铃对冲成就的。

懂得在生活中使用对冲法则的人从不会焦虑或者为任何事情沮丧，无论是失恋还是创业失败，因为你永远不知道上帝明天为你准备了什么，只要你自己不对生活说 No！

这也是我认为丰富广博比专于一处更好的原因，因为要想利用杠铃法则，实际上需要动用的是你的全部知识、爱好和所有的积累。如果你只有一种技术，那么你就是个一头重的杠铃，只能在自由落体中显身了。

最后引用《反脆弱》中的第一句话：风会熄灭蜡烛，却能使火越烧越旺。

归零思维

有一个一般的规律，人拥有得越多就越害怕失去，从而进入一种守成心态，而这正是人的脆弱性。拥有得越多，在风险中可以丢失的东西往往会多于随机中的获益，正如一艘满载财宝在风雨飘摇中的船。而相反的情况则是，当你所处的位置较低时，向下空间有限而向上空间很大，波动性造成的平均收益为正。想想那些一无所有的人，往往无所畏惧——奴隶可以丢掉的只有枷锁。当然这不是说都去当乞丐，起码要把自己的心态归零，是有帮助的。正如乔布斯说的：Stay Hungry, Stay Foolish。

博览群书

博览群书是最佳的获取反脆弱性的方法。因为恰当地使用杠铃策略直接依赖于你知识的广度，但并不是书呆子似的死读书，而是要恰当地把握一类事物的轮廓和精要，而非死缠烂打。因此最好是每一种学科都有所了解，但并非读那种特别有深度的书。例如，若只是想了解哲学就没有必要读完一整套纯粹理性批判类的书。当然，你的专业除外。

宽以待人，有容乃大

宽以待人，其实也体现了反脆弱性。因为，即使自己不总是占便宜的那个，造成的损失在较长时间来看也并不算什么。但是不宽容别人或斤斤计较所造成的损失，或因此失掉的机会，往往是致命的（尤其在信息时代）。因此宽容和博大的气度，具有很高的反脆弱性。

实用理想主义

杠铃策略之一就是实用主义和理想主义的结合。最优秀的理想主义者，往往要奉行最强大的实用主义原则，理想和实用就是杠铃的两端。

懂得杠铃主义的人会把一些事情用最大的实用主义解决，然后就可以无忧无虑地搞理想。比如大学考试，对于无心学术圈的人，他们如果能够巧妙地把成绩控制在 60 分，就比努力得到 80 分的人聪明，因为一旦大学结束，用 20 分换来的“其他技能”往往会帮他们做成自己想做的事情。

4 从物理角度看复杂

非线性动力学，是用物理学的思维理解复杂系统问题的一座丰碑，也是非常有前途的工具学科，它为大数据时代提供了潜在的分析引擎。

为什么说非线性，因为物理之外的系统大多数不能用线性系统表述（详情请见《动力学是如何做预测的》）。

动力学的核心使命是预测系统的变化，非线性动力学的核心使命也是如此。一个经典的非线性动力学系统具有标准的表述形式：

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x)$$

预测一个系统的未来，你需要知道它在微小时间尺度里的性质并列出动力学方程。

$\boldsymbol{x}$ 是一个向量（vector），它所具有的分量个数即系统的维度。

维度是动力学系统的最基本属性，它决定了系统的复杂性，以及其可能具有的基本性质。还有，我们有多大把握预测系统的未来。

高维空间绝非只存在于宇宙之边（广义相对论）或者加速器的深处（弦论），而是你我的生活中处处皆是。

本篇笔者将从低维到高维的顺序，用图形的思维，讲述复杂性是如何随着维度的升高而产生的。

最简单的系统是一维系统，预测一个一维的非线性系统，往往只需抓住一个关键性信息——定点。

维度，动力学和生活（阅读难度★★）

马尔萨斯人口论合不合理？

18 世纪末，在工业革命前夜的英国，一个叫作马尔萨斯的伟大思想家提出了这样一个困扰了人类几个世纪的问题：人类的人口呈指数增长，而食物的总量至多成代数增长，所以当人口的增长超过食物时，人类将不可避免地陷入饥荒、疾病和战争。而普遍性的贫穷，是人类文明的宿命。

这个理论解释了为什么许多古代文明陷入发展停滞的泥沼，例如埃及。

马尔萨斯的理论，其实诠释的是一个叫作 Fix Point（定点）的动力学概念，即在一个复杂系统里，事物的增长往往不是线性的，而是存在一定的稳恒状态，系统的变化会逐步减速并自发地把自己维持在这个状态上。

这样的现象在生活中不胜枚举。比如说小孩子长高到一定程度就不长了；你在网上发状态，开始有很多人点赞，但在一定时间后减速直至停止。

非线性动力学用定点来描述这种现象。为什么定点普遍存在？因为负反馈的普遍存在。当一个事物向一个方向走得太远时，就往往有一种反方向的作用力把它拉回，有点像我们所说的物极必反或阴阳相抵。

马尔萨斯的人口论符合一个叫作 Logistic Model 的经典一维动力学模型，它也因它那美妙绝伦的 S 曲线而闻名。

这个模型说的是，在没有环境压力的时候（人人吃饱饭）人口的增长率是恒定的，所以如果第一年是 2，那么第 N+1 年即为 2 的 N 次方（几何增长），但是一旦人口接近环境的阈值，就会有人开始饿死，而这个饿死的比例会随着人口的增长而增大（负反馈）。这样，当饿死的人的数量等于出生的人的数量时，两个此消彼长的要素就在

某个点上平衡了，即所谓的定点。

反映在数学上，就是这样一个微分方程：

$$\frac{\mathrm{d}N}{\mathrm{d}t} = rN(\frac{K-N}{K})$$

人口的变化取决于两个相乘的因子，一个描述增长（rN），一个描述饥饿（1-K/N）。定点，就是使微分（人口变化率）为 0 的点，当人口数恰好处在这个点上时，就会不增不减。

这个定点具有一个更深刻的性质，无论人口一开始是多少，只要 K 给定，系统就会趋于一个相同的值。这个值由环境本身的容量所决定。

这个微分方程的解是一条优美的 S 型曲线（Sigmoid Function），如图 4.1 所示。它的身影在自然界中比比皆是，反映了自然生长的一般规律。

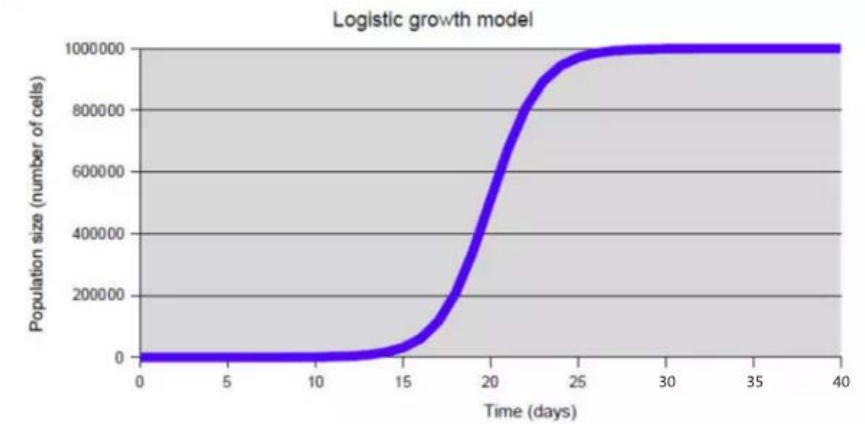


图 4.1

马尔萨斯的确是一个有着深刻洞察力的思想家，他在没有任何这些数学概念的时候发现了这一原理。当人口的增长达到一定限度，大规模的饥荒和战争将使人口增速变慢，从而实现大自然的平衡。

定点的稳定性

动力学里最重要的概念之一是定点，但是定点本身却只有很少的信息，更关键的

性质来自于对定点周围区域的分析，或者说定点的稳定性。

在一些情况里，定点好像是系统变化的宿命。起点是什么并不重要，你不需要担心输在起跑线上，只要你起跑了，就会到一个地方——定点。而在另一些情况里，定点虽然存在，但是只有在极特殊的条件下才能达到，类似于逆袭。逆袭是有的，但是要有极好的运气和相当高的智慧才行。即使你达到了这样的定点，稍有风吹草动也会失去它。

我们用一个被称作稳定性的概念来描述这一特性。稳定性描述的是系统处在定点周边的状态，它是比较容易进到定点还是离开。

一个典型的例子是单摆，单摆的微分方程有两个取零的点，但是你通常看到摆处在最低点却极少有机会看到一个处在顶点的单摆。原因很简单，单摆的低谷是稳定定点，而高点是不稳定的。除非你一开始就静止在最高点而且排除任何外力，否则最轻微的偏离就可以导致单摆回到稳定的最低点。

从物理的角度很容易理解一个定点是稳定的还是不稳定的，只需要稍微离开定点，看一下系统的运动情况，看看系统在定点的相邻区域里的运动趋势怎么随位置变化的。而这翻译成动力学语言就是在定点周围进行泰勒展开，并取一阶线性近似（在一维得到一个线性的斜率，高维就是雅可比矩阵的特征值）。如果在定点周围的运动趋势指向定点（线性的斜率为负，雅可比矩阵特征值为负），则定点在局域内稳定，反之则不稳定，如图 4.2 所示。

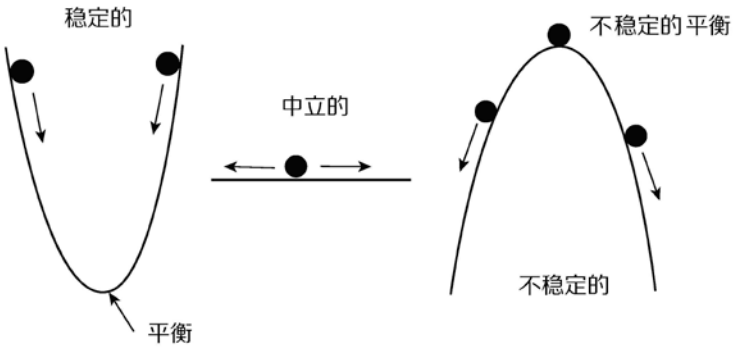


图 4.2

注：定点的稳定性，取决于泰勒展开的不为零的第一项的正负。图 4.2 的左图为稳定平衡，右图为不稳定平衡，虽然均为定点，但周边性质迥异。

稳定性，换一个词叫吸引力。一个稳定性定点，就像一个区域的主人，它能把进入其辖区内的所有人都吸收到它的点上。它所管辖的区域，称为 Basin of Attraction。它是强韧性的代表，无论你怎么干扰它，迫害它，结局都将归于它。找到 Basin of Attraction 是利用定点预测系统的必备条件，给定一个系统，如果它的初始位置处在 Basin of Attraction，那么它必归于定点。

不稳定性呢，就是脆弱性的代表了，任何环境的风吹草动都能结束它表面的美丽。

最强的定点具有全局稳定性，即无论任何初始条件，系统都将趋于这样的定点，这样的系统就是高度可预测系统。

很多系统往往是一个稳定点和一个不稳定点成对出现。比如刚才的人口模型，人口为 0 就是一个不稳定平衡点。当人口为 0 的时候，它可以永远为 0，但只要系统的人口增长了 1，它就会趋于定点 K，掌控系统除 0 之外所有区域的稳定点。

判断简单系统，抓住定点就是抓住了命门。

二维系统与振动

请先看下面几个问题：

为什么振动普遍存在？

为什么自由竞争的结果往往是垄断？

如何理解经济周期的运行？

解决这些非常基本的问题，我们需要一个二维的动力学系统。二维可以描述比一维丰富得多的现象，正如同物理学从描述两个物体的相互作用开始描述了世界。

一维的系统往往归于单调的定点，而二维系统的主角却是振动，也是人类几千年来描述自然最有利的工具。

看看我们周围，从自然到人类，世界可以看作一部不同频率振动组成的交响曲。

四季周而复始，太阳升起落下，我们的呼吸、脉搏、心跳、新陈代谢，生命的更替，经济系统的周期涨落等，几乎有运动的地方，就有振动。

为什么振动的形式如此广泛地存在？其实是因为定点的广泛存在。振动就是围绕一个确定状态的上下波动。就好像希腊神话里痛苦的西西弗斯，把石头推上山，可它却滚下去，然后他又推上山，他想让石头停在山上不动，但却做不到。

为什么振动如此普遍，非线性动力学之父庞加莱给出了一个神一样的定理：Poincare-Bendixson Theorem。

条件如下。

（1）2D：你有一个二维的动力学系统。

（2）Continous：系统连续可微。

（3）Confined：动力学流在一个区域内封闭。

（4）No Fix Point：在此区域内定点不可达到。

结论：

该区域内的动力学流将收敛于一条闭合轨道（等价于圆）。

翻译一下，相平面的闭合轨道=周期性运动=振动。这个定理告诉我们，有限二维系统里的运动形式只有两种：（1）平衡态（归于定点），（2）周期运动。不存在其他情况。有限指的是系统不会无限取值或发散。由于自然中负反馈的普遍存在，因而这一条一般是满足的。这条定律解释了振动普遍存在的根本原因，因为它是二维运动的范式。

作为一条以拓扑学为根据的定理，它标志了人类思维的新形式，即拓扑思维。这种把各种不同形式的系统归于空间里的拓扑研究的思想，是一种超越性的思想。它标志了数学在解释世界的能力上的新高度。从此，我们对世界的认知，取决于我们对几何空间的拓扑性的归类。那些能够归于同一拓扑结构的系统，都具有相同的动力学本质，即使它们的物质组成各不相同。因此，拓扑的思维具有高屋建瓴，以一敌百的特性。

庞加莱定理告诉我们，二维动力学流不是流向定点，就会形成闭合轨道。

这条定理确立了非随机的二维系统的绝对可预测性，二维系统没有混沌。

当你发现振动，你就去找系统里有没有两个关键性的动力学变量，并且观察这个系统是否有稳定的平衡态（如果没有，则往往预示存在一个无定点的闭合区域），这样的方法非常有效。

能量守恒系统的振动

经典物理的振动核心在于能量守恒，无论是弹簧的振动、单摆，还是电磁波。

对这类系统的传统解法是对微分方程进行积分得到运动的轨迹，但是利用一些动力学的基本知识我们也可以完全不用积分就能了解它的运动性质，如图 4.3 所示。

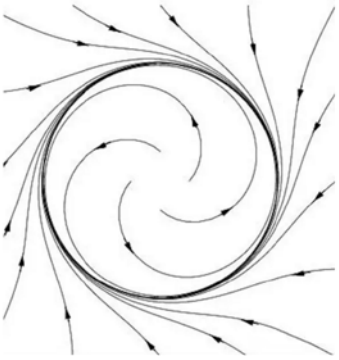


图 4.3

下面以单摆为例。首先，把微分方程写成标准的二维动力学系统形式，即体系内有两个动力学变量：角度和速度。

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = v$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\sin\theta$$

然后，寻找定点（动力学流为 0），显然是θ=0，v=0。

看看定点可不可以达到。只要能量不为 0 的系统都不可以，因为θ=0，v=0，意

味着系统能量为 0。

根据能量守恒定律，能量不为 0 的系统无法到达这个点。

最后，系统是否在相平面里封闭？是的，回复力-sin（θ）起负反馈的作用，根据能量守恒定律，θ和 v 均在有限区间取值。

因此，系统在相平面内做圆周运动，如图 4.4 所示。

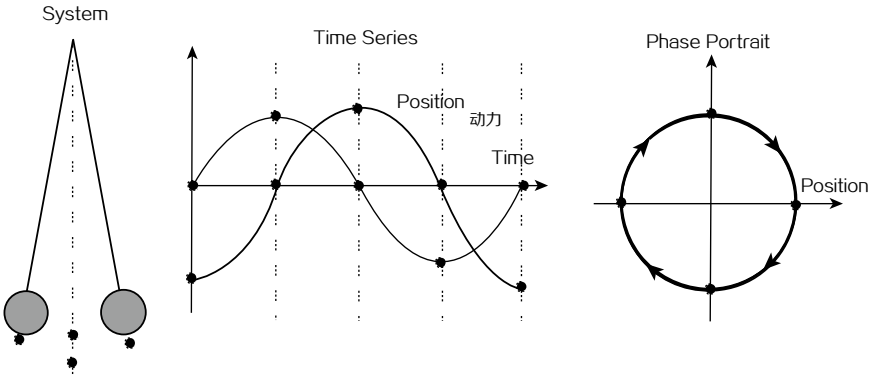


图 4.4

单摆在相平面的圆周运动（右）对应其在真实空间的振动（左）。只要把握了相空间的性质，无须解方程也可以了解运动的性质。

注：如果系统内有摩擦力，能量守恒不再成立，则系统具有稳定点。x=0，θ=0，依然符合庞加莱定理。系统如果不再振动，则归于定点。

当然，经典物理的例子基本是不太重要的，动力学的威力要在更复杂抽象的系统里显现。

能量不守恒的开放系统里的振动

当一个系统有能量流入流出时，我们称之为开放系统（为与能量守恒的保守系统区分），对于一个二维的开放系统，庞加莱定理依然成立，系统若不归于平衡，则步入永恒的循环。

这样的系统比比皆是，神经细胞周期性的电震荡、心脏的跳动（心肌细胞的电震荡引致）、宏观经济运动的周期等。这类问题，显然具有比两个多得多的变量，但是如果把关注点集中在它们进行周期运动的时间范围，往往可以抓住两个关键性变量，从而用二维动力学系统的知识来解决。这体现了物理的核心思维：Reductionism（简化，抓住主要矛盾）的应用。

例如，在经济周期的问题里，两个关键性的变量是国民收入与资本存量。对于神经细胞，关键性的变量是电位和疲惫指数（放电导致的疲惫）。

注：黑格尔的辩证法说，正是矛盾导致变化。翻译过来，就是当系统有两个相互制约的变量时，就会引起永恒的振动。抑或说，Poincare-Bendixon 原理就是辩证法的数学精确表达。

一个最典型的例子依然是延续我们关于物种数量的故事，刚才讲到一维的人口模型里人口将达到定值，而事实上，自然界中的物种数量却是震荡变化的，为什么？

要想解答这个问题，就需要讨论两个物种共存的情况（二维）。试问下面的问题，在一片草原里生活着狮子和羚羊，狮子吃羚羊，羚羊吃草（假设无限），假设一开始物种数量是均等的，那么后来两个物种的数量变化会是怎样的呢？显然，两种物种间有相互作用，狮子的存在依赖于羊（简单的想法是羊肉变成了狮子），而羊的数量因为狮子而减少。如果没有狮子，羊的数量增长就符合之前的 S 曲线。

这个系统可以用一个 Lotka-Volterra 方程的经典二维动力学系统表述:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = ax - \beta xy$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = \delta xy - \gamma y$$

这个方程极为容易理解。系统的两个变量一个是羊的数量（x），另一个是狮子的数量（y）。第一项描述羊的自然生长率。第二项描述羊被吃的数量，x 和 y 的乘积决定了两个物种相遇的机会，所以羊被吃的速率正比于 xy。相应的，狮子可以理解为由羊肉转化出来的，所以其增长率正比于捕获的羊数量（方程二第一项），方程

二最后一项描述狮子的死亡率。

那么，如何预测两种物种数量变化？首先进入相平面，我们看到系统的流形（每一点的微分（dx，dy）构成一个向量，画出箭头犹如流体力学的流速线）。然后我们分析定点，二维系统里含有两个微分方程，如果一个微分方程为 0，例如 dx=0，我们将得到一个代数关系 $x=k\cdot xy$. 在相平面里这对应一条线，即 Nullcline。在这条线上，第一个变量处于平衡态。同样的，我们可以找到变量 y 的 Nullcline，对应相平面的另一条线，这两条线如果有交点，即二维系统的定点，或者说系统的平衡态。

但本章讨论的是稳定性而非平衡态本身。

这个问题可以很容易找到四条 Nullcline 和两个定点：一个是（0,0），另一个是第一象限中的（a,b）。（0,0）代表两个物种都灭绝了，这种情况是羊死光了才可能出现。假设狮子死光了，那么羊就会无限增长（远离定点）。

在相平面上，就表现为动力学流沿着 y 轴（对应羊死光的情况）收敛为 0，而沿着 x 轴（对应狮子死光的情况）发散，如图 4.5 所示。

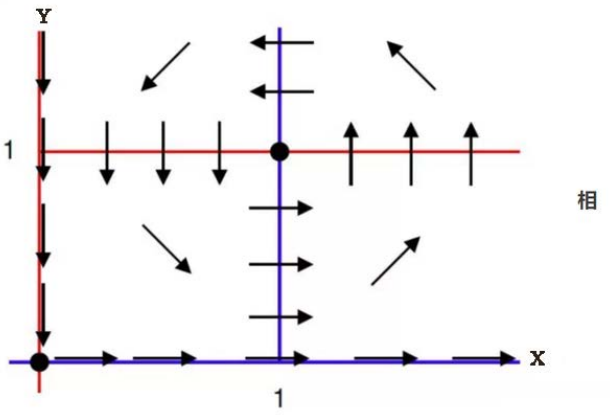


图 4.5

这一现象的隐含含义是（0,0）点在 x 方向上是不稳定性定点，而在 y 方向上是稳定性定点。这种在一定方向上收敛，而在另一些方向上发散的定点，被称为 Saddle Point（鞍点），因为三维空间的势能曲面形如马鞍而得名，如图 4.6 所示。

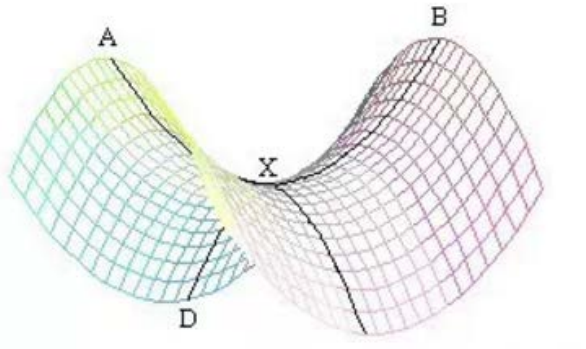


图 4.6

再看看第一象限内的定点（a,b），它描述两个物种数量互相制约的平衡状态，看似是一个合理的结局。当狮子和羊的数量达到平衡时，这不就是生态平衡吗？如果这样想，那么就是停留在初中生物课本了。在这个定点周围找几个点，通过画出（dx,dy）的箭头即可知道，它们都不是朝向这一点，而是围着这点转圈。利用庞加莱大法可知，系统将永不能陷入这个点。而是围绕这个点形成闭合轨道，即振荡。系统的两个物种的初始数量只要不是有一个灭绝或恰好开始就匹配平衡，就都将形成一个振动变化关系。

狮子和羊在固定系统里的数量里周期震荡。图 4.7 中上图为相平面，下图为两个物种数量随时间的变化关系。

整个生态学可以用动力学语言描述。其核心议题，即生态系统的稳定性，正是动力学最擅长分析的内容。

Lotka-Volterra 系统在经济学中也有重要应用。凯恩斯学派用以解释劳动雇用率和资本的周期震荡。这一理论把资本对应为狮子，而劳动雇用率是猎物，两者总是不能自发地处于定点（100%雇用率），而是进入周而复始的震荡状态。

甚至整个凯恩斯的理论可以放入一个简化的二维动力学系统。生产和需求作为一对互相追捕却永远捕不到对方的对手，将陷入不停歇的振动状态，即经济周期。它导致经济运行不可避免的在一定时间走向低谷。

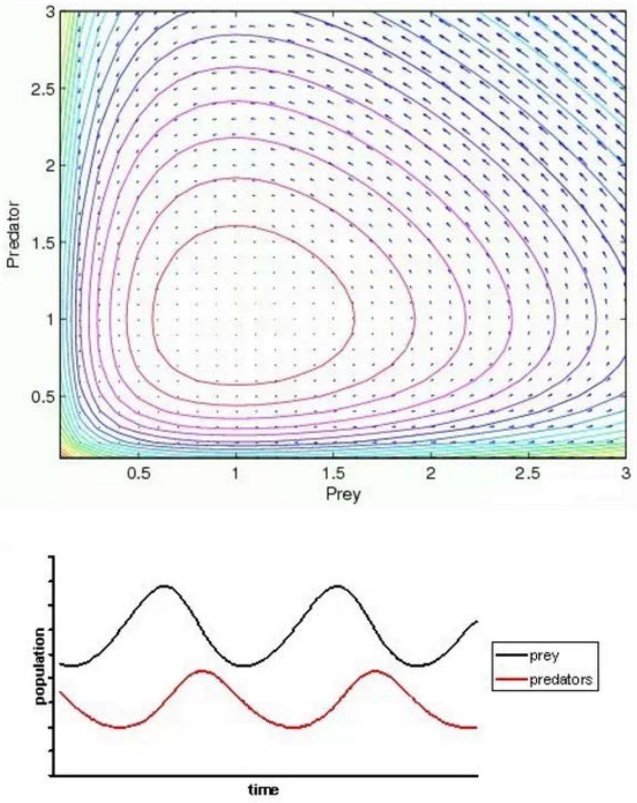


图 4.7

但是凯恩斯的理论基础终究是松动的，因为经济系统不是二维的，过多简化的理论可以帮助我们理解现象，如果以此为实践的基础，则是危险的。

在动力学的世界里，无所谓色彩味道，只在乎系统所包含的自由度，或者说维度，一定的维度对应一定的现象。

维度是动力学系统的最基本属性，它决定了系统的复杂性，以及其可能具有的基本性质。还有，我们有多大把握预测系统的未来。高维空间的属性绝非只有爱因斯坦在乎，而是与你我息息相关。

一维的世界里我们看到的是反馈和定点，或者说静态平衡的法则。二维的世界里我们看到的是振动这一自然永恒的主题，也是大自然的动态平衡。下面将继续对二维系统进行讨论，并引入三维系统和那里才会出现的混沌，最后直指高维系统和复杂网络。

二维系统可以稳定存在的运动状态有两个，一个是定点，另一个是振动。这两种状态可以转化，有的是从一个定点到另一个定点的转化，也有的是定点和振动之间的转化，这就涉及一个非线性动力学永恒的主题——Bifurcation（分叉）。

分叉与相变

我们用分叉研究一个动力学系统的演变。动力学系统由状态变量（系统可以自由变化的量）和控制变量（参数）组成。在初步讨论一个动力学系统性质的时候，我们先假设参数不变，因此可以得到系统动力学在相平面的拓扑图，然后求出定点和轨道。在二维的世界里，参数给定后，即可得出动力学流型，从而一切皆可精确预测。

真实的世界里从来没有一成不变的参数，真正不变的只有变化，甚至有的时候参数和变量甚至难以区分。因此，非线性动力学给出了对世界的最精密的描述，不是确定参数下的流型，而是在参数空间里对应的不同相平面的流型。简单地讲，动力学不仅感兴趣我们现在所在的世界，而是所有可能的世界（每个参数就是一个世界）。参数的空间好比小径分叉的花园（无限可能性的博物馆），每一点上你都有一扇窗户，打开可以看到那个世界的可能性。在这个花园里，你将看到一种可能性是如何演化成另一种可能性的。

下面我们就到二维世界里去转一转，请看图 4.8，可以看到十字线和一个抛物线。这是最简单的线性二维动力学系统，完全可以通过求解系数矩阵（对应于一维情况下的单一系数）的特征值解决。首先看这个系统的定点（0,0）（带入方程微分为 0），系统是被这个定点牢牢抓住，还是围绕它振动，还是远离它而去，则取决于系统的参数。

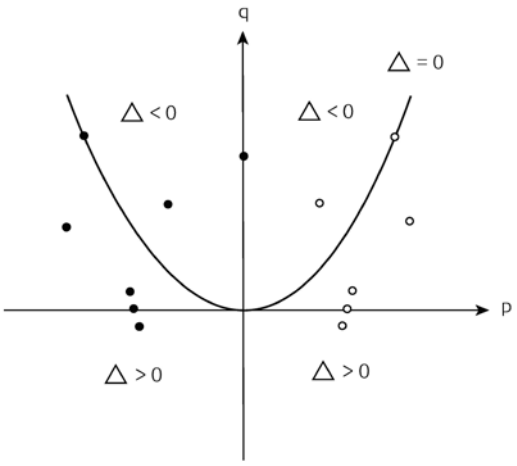


图 4.8

这个平面的横轴和纵轴代表了这一矩阵特征值的实部和虚部。当系统的参数变化时，表现为系数矩阵的特征值在这一平面上的运动。特征根的实数部分的正负决定了系统是趋于稳定的定点还是发散。若是为负你将收敛到到它身边，为正则你将远离它而去。而如果实数部分为 0，特征根只有虚部，那么意味着系统既要远离定点又跳出不去它的引力范围，最后就变成围着它转，即振动的情况。虚部的正负决定了系统围绕定点转动的方向，在此不多叙述。

那么什么是 Bifurcation 呢？它就是参数空间里系统动力学流的性质发生质变的点。例如，图 4.8 中的那个抛物线，当系统的参数变化越过抛物线时，系统就从稳定吸引变成了发散远离定点，这个过程就是 Bifurcation。

而在抛物线一侧的变化只是定量的变化，却无定性改变，这就是普通的变化。Bifurcation 标志了系统的动力学性质发生彻底的变化。好比两个人在一条路上走着走着，突然到了岔路口，从此南辕北辙。

在动力学家的眼里，只有那个 Bifurcation Point 具有关键意义，可以起到区分不同系统的作用，其他小的变化都忽略了。

另一种典型的 Bifurcation 情况如图 4.9 所示。

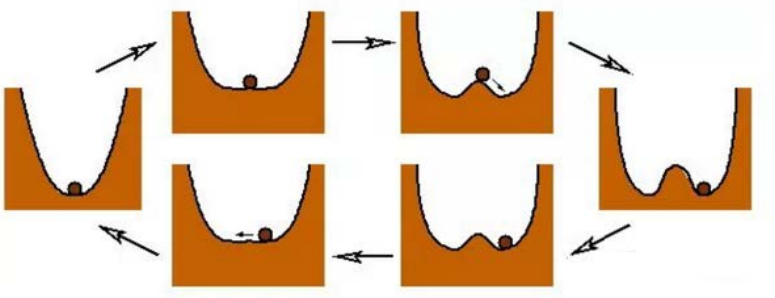


图 4.9

图 4.9 中的小球一开始在谷底，处于稳定平衡。这个谷就代表系统的参数，当参数固定时，山谷的形状就是确定的。当我们改变参数时，山谷的形状就会发生变化，谷底逐渐被拉平，最后隆起出一个小山。在这个过程中，中心点的稳定性丧失，小球将面临一次全新的选择，是向左还是向右？这个谷底变成小山的过程就叫 Bifurcation，而谷和丘的临界点，就是 Bifurcation Point。

从图 4.9 可见，Bifurcation 的本质是系统反馈性质的变化。当小球在谷底时，一个负反馈保证它不离开（稳），而当谷底逐步变平凸起的过程，负反馈演化成离开谷底的正反馈。

Bifurcation Point 上的小球具有“自由意志”，或者说非常敏感，一个随机的扰动都可以被放大（正反馈的作用），使它向左或向右。这就是历史的转折点。而当 Bifurcation 的过程结束时，小球就落入了新的平衡点。此时的它，已经被一个负反馈束缚住，除非有强大的能量，否则不会离去。

Bifurcation，正是物理里相变的化身。在动力学的世界观里，那些定量的改变等于没变，而只有 Bifurcation——分道扬镳，才是真正的变化。物理、化学、生物一切最有趣的现象，都在 Bifurcation Point 上，因为它的敏感，成就了它的无限可能。

> 注：Bifurcation Point，就是我们所说的决定性瞬间。在这个时候，系统的前途未卜，而任何一个风吹草动都可能使它转左或转右，从而走向截然不同的未来。如同高考考场上的同学，蒙对或蒙错一个选择题可能就去往了截然不同的城市，遇到了截然不同的爱情。

Hopf-Bifurcation：沟通平衡与振动的世界

用一句话说，Hopf-Bifurcation 描述的是一个系统定点失去吸引力并最终产生闭合轨道的过程。这与图 4.8 其实是一回事，我们把非线性系统在定点附近进行线性近似就可以沿用上面的分析。

BZ 反应（Belousov Zhabotinsky 化学反应）

在我们高中课本中有个东西叫化学平衡，说的是化学过程最终都将导致平衡，该反应的反应过了，我们最终会得到一堆万年不变的反应产物。但是 20 世纪 50 年代，一个苏联科学家 Belousov 却在他的反应里发现了一个十分惊人的现象，他发现他手里的混合物反应后还会在一段时间回到原来的状态，然后又重新反应，如此周期反复。这一现象一出，他就被封杀了。因为他的结果不符合热力学第二定律（根据热力学第二定律，自发状态下系统必须趋于平衡），又加上适逢冷战，他到死也没看到他的成果被承认，成为科学史上重大悲剧之一。

但是他的发现却开拓了一个全新的领域：化学震荡，而他的发现也成为复杂性可以从简单系统中诞生的典型例子，与图灵对生物斑图的研究一起，开拓了复杂科学的先河。

Belousov 的化学震荡可以自发产生美丽复杂的斑图（见图 4.10），被认为是复杂性从简单系统产生的典范，对生命起源等问题都很有启发。

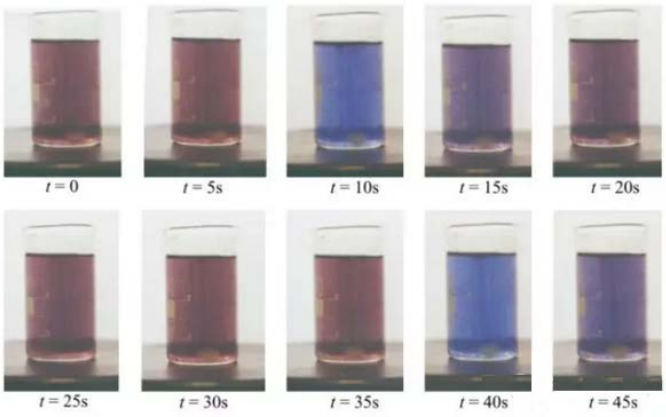


图 4.10

周期震荡的化学反应，红变蓝又变红。

如果我们给这个化学反应写出热力学方程，就可以发现，循规蹈矩的化学平衡和“异常”的化学震荡可以完全统一在一个系统里，只是根据反应物浓度不同而不同，它的本质即 Hopf Bifurcation。

Belousov 反应具有众多反应物和接近 20 个步骤，但是可以简化为一个二维动力学系统（内容繁杂在此不叙）：

$$\dot{x} = a - x - \frac{4xy}{1+x^2}$$

$$\dot{y} = bx\left(1 - \frac{y}{1+x^2}\right)$$

随着参数 a，b 的变化，系统具有完全不同的动力学模型，如图 4.11 所示。

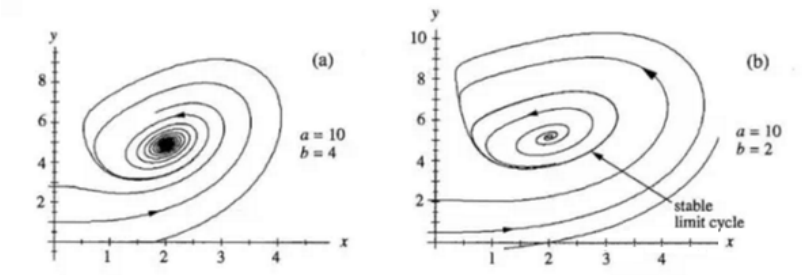


图 4.11

图 4.11 中，左图是一个具有静止平衡态（定点）的系统，动力学流从不同的位置旋入这个系统。右图为振动解（Limit Cycle）的诞生，事实上，两张图描述的是一个系统的连续变化，开始那个稳定的平衡点失去稳定属性，流型从旋入这个点变为旋出，从而归于确定的闭合轨道。这就是 Hopf Bifurcation 的范式。

Hopf Bifurcation 作为阐述振动和静态平衡互相演化的基本手段，在生物、经济等领域反复出现。

甚至我们的生命过程本身也可以理解为一个大的 Hopf Bifurcation。心脏的跳动和新陈代谢的循环伴随我们一生，这是系统的振动解。我们死的那一刻，振动停止，

我们步入了静态平衡。这就是 Bifurcation Point，from live to death。

高维系统与混沌

当系统的维数达到三维时，主宰动力学模型的就不再是那些稳定可测的点或圆环，而是初值敏感，极难预测的混沌。

混沌其实没有你想的混沌，系统依然具有确定性的方程，只是其复杂性使得它看上去像是随机、毫无秩序而已，所以，混沌实则乱中有序。

笔者用一种最简洁的方法说明混沌是如何产生的。在一维系统中我们反复强调定点，因为定点是一维系统唯一可以具备的稳定状态。而在二维非线性系统则反复强调闭合轨道，因为定点和闭合轨道是二维系统唯一可能的稳定状态。那么三维非线性系统，可能的稳定状态是什么呢？遵守点线面的顺序，你一定猜到了是曲面。对，三维系统的稳字状态是三维空间里复杂的曲面。只是说一个曲面稳定已经不再有意思了，我们管它叫吸引子，是三维空间里吸引系统进入的一个物体。

图 4.12 为吸引子去曲面的范例——洛伦兹吸引子的形状。

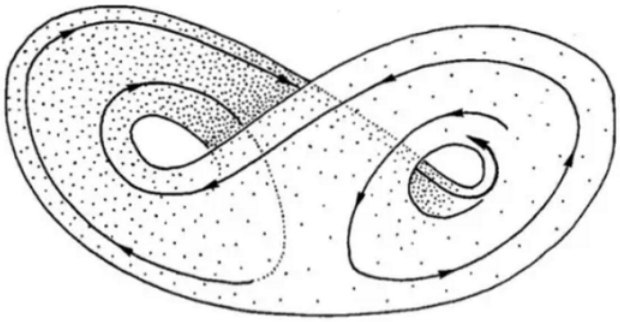


图 4.12

那么为什么三维非线性系统可以产生混沌？因为物体被一整个曲面吸引，不知道往哪里去了。即使它被紧闭在这个曲面上，也可以具备无数的轨道（面上的曲线）。轨道变得复杂不可预测，因而混沌。

洛伦兹以它优美绝伦的洛伦兹方程证明了混沌是如何从一个三维的确定性系统里产生出来的。洛伦兹方程产生的混沌，被后人称为蝴蝶效应。传说南美洲的蝴蝶煽

煽翅膀，就可以引起北美的飓风。即使这是真的，它也不是洛伦兹的本意，因为蝴蝶效应说的是，动力学流动在相空间的姿态类似一只蝴蝶！

洛伦兹方程具有优美简洁的形式：

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sigma(y - x)$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = \gamma y - y - xz$$
$$\frac{\mathrm{d}z}{\mathrm{d}t} = xy - bz$$

当这组方程的参数正好位于 Hopf Bifurcation 的点时，我们就能得到翩翩起舞的蝴蝶。

图 4.13 为洛伦兹吸引子，由洛伦兹方程确定的三维系统具有两个吸引中心（定点），系统围绕两个定点旋转，形成极为复杂不可捉摸的轨道，形如扇扇翅膀的蝴蝶。

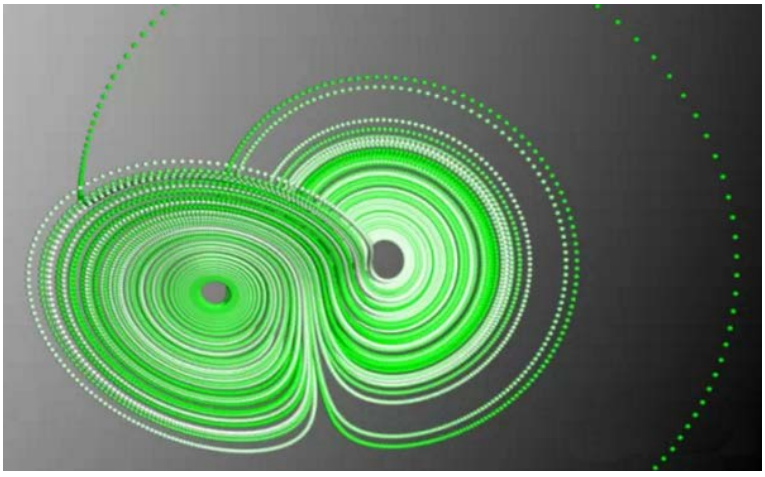


图 4.13

对这一系统最简单的理解是，洛伦兹系统依然描述闭合轨道，这和之前描述的二维系统的振动是相同的。但是在三维系统里，我们有两个定点，以及随之确定的吸引子曲面。系统时而围绕着其中一个定点旋转，时而围绕着另一个定点旋转，但是什么时候改变围绕旋转的对象却是不可测的。

混沌准确的定义是相邻轨道的稳定性，或者说初值敏感性，你是否可以从一步之遥，发展到天壤之别。不过和之前分析定点的稳定性不同，这次需要分析的是轨道的稳定性。

如何分析轨道的稳定性？很简单。我们选取在相空间里起初无限接近的两条轨道，分析它们之间的距离在之后是扩大还是缩小。如果两条轨道的距离是扩大的，则预示着系统是混沌的。

我们利用一个叫 Liyaponov 指数的量来分析轨道距离变化的趋势。如果指数为正，则意味着开始一步之遥的两条轨道会变成天壤之别；而指数为负，则意味着它们将归于一处。

图 4.14 为混沌的判定原理：初值敏感性，无限靠近的两条曲线，未来会怎样？

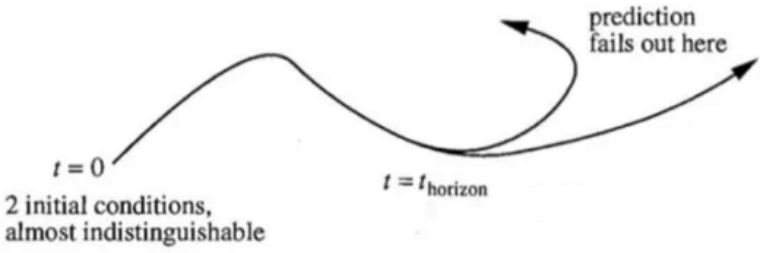


图 4.14

混沌与秩序

混沌实则是复杂秩序的产生者，它所产生的秩序，叫作分型结构（Fractal）。分型结构的本质是自相似性，或者说标度不变形。也就是说，把它放大或缩小 N 倍仍和原先长得一样，或者说宇宙里包含着小宇宙的无限迭代形式。分型是自然界中图案的主宰，从树叶到海岸线，甚至我们的肺都具有此类结构。分型如此常见，是自然界中的混沌动力学体系写下的诗篇。每一个分型结构的背后，大概都藏着如蝴蝶翩翩起舞般美丽的动力学方程。

图 4.15 为混沌体系产生的某种图案，像不像叶子？

图 4.15

混沌与市场自由

混沌是美丽的，因为它代表自由，自由竞争的市场往往最后产生垄断（定点），但垄断格局却从不持久，因为参与市场竞争的个体实则无数，所以市场其实是个高维的混沌系统，而对于这样的系统，即使一时产生垄断，其风云莫测的性质也会打破它。

> 注：混沌的不可预测其实是描绘初值敏感，两个起初靠在一起的轨道注定要发散。但是它终究是确定性系统，与量子力学的不确定不同。

复杂性摧毁了人类预测一切的伟大梦想，同时，我们却在另一个方向上捷报频传，这就是产生我们目前人工智能基础的机器学习。突然发现，可以用数据驱动的机器学习帮助我们挡掉一些复杂性。

第二部分　机器学习

5 白话机器学习

（阅读难度★）

机器学习是什么

一段程序可以看作一连串从输入到输出的过程，无论是工程师还是程序员，我们都想通过设计来完成某种功能。以做一个网页为例，要画视觉图、UI 图，以及前端后端交互图等，我们要给计算机设计一套解决具体问题的流程。

而机器学习呢？机器学习是你不去设计，而是让计算机自己去解决，如同用一套模子去打磨出能够解决特定问题的武器。机器学习做的正是“自发能够产生解决问题的程序”的程序。一些机器学习的经典算法，如线性回归、SVM、神经网络等，单个都不能解决问题，但是通过“学习”，却可以预测房价或寻找美女图片。

我们可以这样理解机器学习：把它倒过来变成“学习机器”，即一个能自动从数据中学到程序的机器。

它是怎么做到的

让机器模仿人学习的过程来学习些程序，最常见的方法就是有老师告诉你对错的学习，这称为有监督学习。

从错误中学习，即监督学习，你可以想象构建这样一个机器人，需要几个部分：

（1）输入系统，录入数据；

（2）思考预测，得到一个结果；

（3）感知错误，记录错误；

（4）根据错误改变思考，避免再犯同样的错误。

下面看看在具体的环境下我们是如何构建这个机器的。

机器学习的老师是人类的婴儿，婴儿开始什么都不会，通过观察别的孩子或家长，学走路，学说话。站不起来，就摔倒；说不对，就挨骂。他（她）知道错误后，知道换个姿势，换种说法。它的眼睛和触觉在获取数据，每一次决定怎么爬起来怎么走都是在思考和做出预测，最后把摔倒和行为联系起来，改进行为就是学习。

举个具体例子——预测房产的价格。一个房子的价格取决于非常多的相关要素，比如房子中卧室、厕所、客厅的数量和面积；周围环境，如交通状况、安静与否、有无学校等，我们需要从这些要素中知道房子的价格。如果你是传统的房产专家，那么你的方法可能是到各个地点调查成交价，随着经验的增加，你会得到各个区域的基准价格，然后按照这个基准价格结合房子的特征上下调整，得出经验公式。

有了最原始的机器学习，我们也可以把这个人类得出经验的过程让机器解决。现在只需在网上放一份表格，让人们把有关房子的有用无用的信息和价格都填了，做一个巨大的 Excel 表，然后让程序从这些信息中学习价格的模式。

在这里，我们将之前的要素换成一个新词——特征。每一个特征，都用一个数表示，如同线性代数里的坐标基。这些特征是如何决定价格的，我们可以不管，让机器去决定。特征的个数我们通常称为维度，一个问题的维度往往决定其复杂性，以及所使用的方法，这就是复杂系统擅长的范畴，在此不详述。维度本身同时决定了我们可能需求的数据量的多少，高维度意味着我们需要求解问题的信息量也成比例的增长。特征工程就是从大量的信息，比如“我的房子有个大落地窗”这种话里提取出和房价可能相关的属性（Attribute）。

所谓机器学习的模型选择，就是在寻找一个基本学习框架，包含你对预测事物的

非常一般的理解，比如房子的价格可能是很多因素的叠加，学名线性回归：

$$\text{price} = \theta_1 \bullet \text{feet} + \theta_2 * \text{neigh} + \theta_3 \bullet \text{sale} + b$$

我们可以用一个算法，让一个非常初级的模型自动进化，机器学习工程师要做的就是把这个模式的毛坯找出来，再把这个寻找自动算法的算法做出来，做出来后，后续的数据就会把这个毛坯逐渐打造成模型，如图 5.1 所示。

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

图 5.1

这个把模型打磨成形的过程就是学习，或者叫求解参数。这组参数不是让某个特定的房子的价格被最好的拟合，而是让整个数据集都被一组参数涵盖。这组参数就是我们学习的结果，仿佛像人一样拥有预测房价的能力。

就像人一样，机器也是从错误中学习的。因此，在最经典的监督学习里，我们首先要做的是衡量错误的大小，我们用一个叫 Cost Function 的函数来衡量模型预测的结果与真实值的差距，模型的效果越差，这个 Cost Function 的值就越高。

$$\text{Cost} = \frac{\sum_{i=1}^{500} \text{MyGuess}(i) - \text{RealAnswer}(i)^2}{500 \times 2}$$

这个函数告诉我们，每次犯错都是潜在的风险和损失，我们称之为 Cost 函数，可用来衡量当前模型离最终正确模型的距离。

这个 Cost 函数恰恰是由刚刚提到的参数决定的。这些参数的含义是，当你改变某一个特征，比如房屋的面积，房子的价格就会变动，但是变动的方式是什么，你不

知道，你就去随机地扰动这些参数看看得来的价格会如何变化，并且和真实的价格信息比较，可能忽然某个瞬间，你发现所有房子的价格都被一组参数神奇的拟合了。

我们称这组参数为最优，或者说这就是学习的结果。虽然这个 Cost 函数和统计学里的方差很相近，但并非一个意思。

机器学习能否成功，就在于能否找到这个最优位置。大多数时候，我们连这个最优位置是否存在都不知道。

当然，刚刚说的随机方法很不靠谱，实际操作中我们用到的是一个叫作梯度下降的方法，让参数顺着最快速减少预测错误的方向自动调整，如图 5.2 所示。

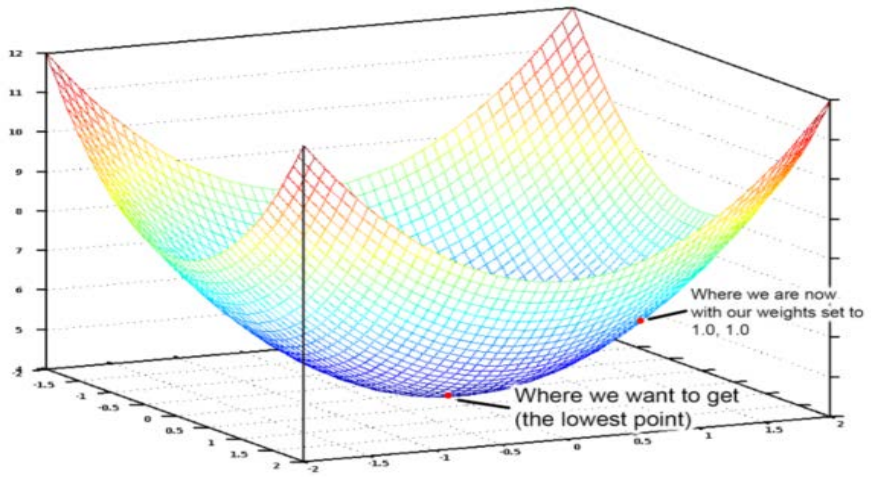


图 5.2

图 5.2 中的 *x* 轴和 *y* 轴代表参数，高度代表错误率，洼地的那个点正是最优参数点。

训练与测试

用一个形象的比喻理解有监督学习，它就像一个拿着一堆模拟考试题学习的学生，这个过程中通过不停地比较和标准答案的差距来完成试卷，而最终这个学生要面对的是真实的考试。相对应的，我们通常在有监督学习里把我们的数据分成两部分，一部分做测试用，另一部分做考试用。一个学生很可能在平时练习的时候把每道考题都背得滚瓜烂熟，而在真正的考试里考得一塌糊涂，这种情形我们称之为过拟合。

过拟合发生的原因，主要就是在局限的数据里过度地挖掘模式导致的。假设一个人想要识别树叶，他只见过那些有锯齿的树叶，因而见到没有锯齿的叶子时就会判断它不是叶子。或者一个人只见过白天鹅，他根据已有的经验去定义天鹅，而当第一个黑天鹅出现时，他的定义就失效了。

无监督学习

在一些更困难的问题里，我们根本无法掌握所要求解的信息，比如房屋的价格，这好比一个没有标准答案的开放性考题，我们需要发散式地探索。

无监督学习看起来不能给出一个准确的答案，但却可以让我们得到意想不到的惊喜。

无监督学习最典型的范例就是聚类，一个典型的例子是，你是一家互联网公司，有一大堆客户信息，首先你需要对这些数据进行整理归类。利用无监督学习里的聚类算法如 k-means，可以轻松地把你的客户信息根据特征自动划分成大类，之后你可能发现喜欢爱马仕包的女生也爱喝拉菲红酒这种信息。

再比如说你有大量的心电图数据，你要分辨出哪些是异常的可能是心脏病的情况，但是你又不知道与病人有关的其他信息，让机器进行无监督学习，很可能得到意想不到的结果。

机器学习很美吗？

机器学习这件事，看上去很美，事实上一点也不，因为你面临的是变幻莫测的问题，你将模型训练得再好，你的预测也是有风险的。首先，因为存在模式，问题才可以预测，如果你的问题本身就没有模式，或者模式极为不可捉摸，那么你的算法再强大也会失败。比如说，你要预测某个商品网站上衣服的价格，但是这个网站里的衣服都是一个疯狂的销售者根据他每天的心情好坏制定的，而你却去找了一堆衣服的尺码、材质特征来，那么你的算法终究要失败。

即使找到了模式，也总会在某个时间点上，碰到问题的边界。比如说，你要做一个根据女生特征匹配对象的机器学习算法，你的算法总是根据你过去掌握的男女匹配记录做的，而总会在某个时刻，你的对象已经不符合历史数据的规律，这时你要么调

整算法，要么直接放弃服务这类无法预测的人。

机器学习最难的是什么？

如果你思考一遍上述流程，可能发现一切都很容易，从数据清洗、特征提取到模型选择，事实上，机器学习最难的一步，这里根本没有提到，那就是把现实生产生活中的问题，提炼成一个机器学习问题。这需要的是你对问题本身的深刻洞察。有一天，也许整个数据清洗到模型选择和交叉验证都自动化了，但始终有一个东西不能完全被机器搞定，那就是你如何从一个全新的领域，提取机器学习可以帮助人类解决的最重要的问题。再有，无论机器的预测多准确，如果它的结果不是在解答人的需求，那么就是一个没用的或者至少不惹人喜欢的东西。

如果机器学习想要解决一个问题，则必须具备三个条件：

（1）系统中可能存在模式；

（2）这种模式不是一般解析手段可以猜测到的；

（3）数据可以获取。

有一个条件不符合，都很难运用机器学习。

机器学习的一个核心任务是模式识别，也可以看出它和刚才讲的复杂系统提到的模式的关系。我们讲复杂系统难以通过其成分的分析对整体进行预测，然而由于复杂系统通常存在模式，因此可以通过模式识别来对系统进行归类，并预测各种可能的结果。

你要让机器做决策，一个基本的思路是从之前统计的数据，挖掘已有的模式（Pattern）入手，来掌握新的数据中蕴含的信息。这个模式在有监督学习的例子里，就是把某种数据结构和假设结论关联起来的过程，我们通常用条件概率描述。那么，让机器做决策，就是不停地通过新数据来调整这个数据结构（特征）与假设结果对应的条件概率。通常我们要把预先对某领域的知识作为预设（Prior），它是一个假设结果在数据收集前的概率密度函数，通过收集数据我们得到调整后的假设结果的概率密度函数，被称为后验概率（Posterior），最终的目标是机器得到的概率密度函数与真实情况最匹配，即 Maximum a Posterior（MAP），这是机器学习的最终目标。

具体来说，机器学习问题事实上可以理解为两个层次。

层次一，建立描述问题的大框架。

这里，你所选择的不是模型，而是一个大框架，或者一个模型空间，它包含与这个大框架自洽的若干假设。下一个问题，是在这样一个模型框架下寻找那个与已知数据最匹配的假设，简称训练过程，或者调参（假设由参数表示）。

层次二，用学到的模型在新的数据里做预测或决策。

对于第一步模型框架的选择，我们要看到不同的模型框架适应的问题范畴不同，并且模型的容纳能力不同，从某种角度看就是描述能力不同。越复杂的模型，描述能力越强（用自身的高维度战胜问题的高维度）。然而，更加复杂的结构同时会带来更高的风险，因此，我们需要建立一个方法平衡描述能力和范化风险，这个方法被称为正则化，通过引入先验来约束模型空间的大小，从而在偏差与误差之间做权衡。

机器学习对抗复杂性的第一个阶段，是尽量寻找那些比较简单的模型，来刻画系统、做决策，因为人们笃信奥卡姆剃刀法则，当一件事情可以用简单的办法简化时，就不要用复杂的工具。

关于模型框架，本书给读者提供一个最小的可用框架，即简单贝叶斯分类器、树模型和神经元。

6 浅谈贝叶斯分析

贝叶斯分析是整个大数据和机器学习背后的思维体系，它的思想之深刻远远超出一般人所认知，我们这里要从贝叶斯统计说起。

首先谈概率，概率这件事每个人都觉得自己很熟悉，但概率的定义，却不一定说的出来。我们中学课本里说概率是用来表述一件事发生的频率，或者说这叫作客观概率。

而贝叶斯框架下的概率理论却从另一个角度为我们展开了答案，这里说概率是我们个人的一个主观概念，表明我们对发生某件事情的相信程度。如同 Pierre Lapalace 说的：“Probability theory is nothing but common sense reduced to calculation。”这正是贝叶斯流派的核心，换句话说，它解决的是来自外部的信息与我们大脑内信念的交互关系。

两种对于概率的解读区分了频率流派和贝叶斯流派。如果你不理解主观概率，就无法理解贝叶斯定律的核心思想。

基于条件概率的贝叶斯定律数学方程极为简单：

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

先验概率在贝叶斯统计中具有重要意义，先验概率即我们在取得证据之前所指定

的概率 $P(A)$，这个值通常是根据我们之前的常识计算的，带有一定的主观色彩。

一个非常有趣的现象是，如果我们的先验概率审定为 1 或 0（即肯定或否定某件事发生），那么无论我们如何增加证据，你仍会得到同样的条件概率（此时 $P(A)=0$ 或 1，$P(A|B)= 0$ 或 1）。这告诉我们的第一个经验就是不要过早下论断，下了论断你的预测也就无法进化了，或者可以称之为信仰。如果你想让你的认知进步，就要给各种假设都留一点空间。

贝叶斯分析的思路对于由证据的积累来推测一个事物发生的概率具有重大作用，它告诉我们当要预测一个事物时，首先需要根据已有的经验和知识推断一个先验概率，然后在新证据不断积累的情况下调整这个概率。通过积累证据来得到一个事件发生概率的过程称为贝叶斯分析。

贝叶斯分析中的三要素（先验概率、条件概率、后验概率）在不同的问题中有不同的侧重点，很多时候我们都会忽略先验概率的作用，比如描述一个人很书呆子气，让你判断他是大学老师还是销售员的经典案例（要看先验大学老师还是销售员哪个多）。但是有时候我们也不理解条件概率，比如著名的辛普森案为了证明辛普森有杀妻之罪，检方说辛普森之前家暴，而辩护律师说，美国有 400 万女性被丈夫或男友打过，而其中有 1432 人被杀，概率很低。这其实就误用了后验概率，这里的条件是被杀而且有家暴，而要推测的事件是凶手是男友（事实上概率高达 90%），这才是贝叶斯分析的正当用法，而辩护律师却在混淆条件与要验证的假设。

理解贝叶斯分析最好的方法即图像法，如图 6.1 所示。这里的 A 的面积即先验，后验是阴影占 B 的面积的百分比。

贝叶斯分析可以瞬间理解一些常用的理论，如幸存者偏差，你发现一些没读过书的人很有钱，事实上你发现的是幸存者（对应图 6.1 中的小红圈），而红圈外的大部分面积你都没见到。还有阴谋论，阴谋论的特点是条件很多很复杂，但是条件一旦成立，结论几乎成立，你一旦考虑了先验，这些条件成立本身即很困难，阴谋论不攻自破。

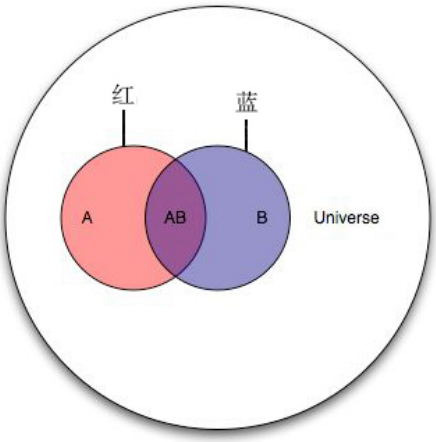


图 6.1

注：图 6.1 上红圈和蓝圈的面积，我们在开始时很难知道，这才是应用中的难点。

此处贝叶斯分析的框架也在教我们如何处理特例与一般常识的规律。如果你太注重特例（即完全不看先验概率），则很有可能误把噪声看作信号，奋不顾身地跳下去。如果恪守先验概率，就成为无视变化而墨守成规的人。那些相信贝叶斯概率框架的人生存率会更高，因为他们会重视特例，但也不忘记书本的经验，根据贝叶斯公式小心地调整信心，甚至会主动设计实验，根据信号判断假设，这就是我们下一步要讲的。

在刚刚讲到的贝叶斯统计分析的基础上，可以引出一个更核心的概念：贝叶斯决策。

贝叶斯决策主要包含四个部分：数据（D）、假设（W）、目标（O）、决策（S）。此处的数据即之前讲到的证据，假设是我们要验证的事实，目标是我们最终要取得优化的量，决策是根据目标得到的最后行为。与上一步贝叶斯分析相比，增加的部分是目标和决策。假设在问题里如果是连续的，则往往以参数空间的形式表达。

然后我们可以按照如下步骤操作：

第一步，厘清因果链条，哪个是假设，哪个是证据。

第二步，给出所有可能的假设，即假设空间。

第三步，给出先验概率。

第四步，根据贝叶斯概率公式求解后验概率，得到假设空间的后验概率分布。

第五步，利用后验概率求解条件期望，得到条件期望最大值对应的行为。

贝叶斯决策一旦变成自动化的计算机算法，它就是机器学习。

7 简单贝叶斯分类器

（阅读难度★）

简单贝叶斯分类器，是贝叶斯分析在机器学习的直接应用硕果，下面就通过故事，讲解一个最简单的分类器——简单贝叶斯。

1787 年 5 月，美国各州（当时为 13 个）代表在费城召开制宪会议；1787 年 9 月，美国的宪法草案被分发到各州进行讨论。一批反对派以“反联邦主义者”为笔名，发表了大量文章对该草案提出批评。宪法起草人之一亚历山大·汉密尔顿着急了，他找到曾任外交国务秘书（即后来的国务卿）的约翰·杰伊，以及纽约市国会议员麦迪逊，一同以普布利乌斯（Publius）的笔名发表文章，向公众解释为什么美国需要一部宪法。他们走笔如飞，通常在一周之内就会发表 3 至 4 篇新的评论。1788 年，他们所写的 85 篇文章结集出版，这就是美国历史上著名的《联邦党人文集》。

《联邦党人文集》出版的时候，汉密尔顿坚持匿名发表，于是，这些文章到底出自谁人之手，成了一桩公案。1810 年，汉密尔顿接受了一个政敌的决斗挑战，但出于基督徒的宗教信仰，他决意不向对方开枪。在决斗之前数日，汉密尔顿自知时日不多，他列出了一份《联邦党人文集》的作者名单。1818 年，麦迪逊又提出了另一份作者名单。这两份名单并不一致。在 85 篇文章中，有 73 篇文章的作者身份较为明确，其余 12 篇存在争议。

1955 年，哈佛大学统计学教授 Fredrick Mosteller 找到芝加哥大学的年轻统计学家 David Wallance，建议他跟自己一起做一个小课题，他想用统计学的方法，鉴定出《联邦党人文集》的作者身份。

但这根本就不是一个小课题。汉密尔顿和麦迪逊都是文章高手，他们的文风非常接近。从已经确定作者身份的那部分文本来看，汉密尔顿写了 9.4 万字，麦迪逊写了 11.4 万字。汉密尔顿每个句子的平均长度是 34.55 字，而麦迪逊的是 34.59 字。就写作风格而论，汉密尔顿和麦迪逊简直就是一对双胞胎。汉密尔顿和麦迪逊写这些文章，用了大约一年的时间，而 Mosteller 和 Wallance 甄别出作者的身份花了 10 多年的时间。

如何分辨两人写作风格的细微差别，并据此判断每篇文章的作者是问题的关键。他们所采用的方法就是以贝叶斯公式为核心的包含两个类别的分类算法。先挑选一些能够反映作者写作风格的词汇，在已经确定了作者的文本中，对这些特征词汇的出现频率进行统计，然后再统计这些词汇在那些不确定作者的文本中的出现频率，从而根据词频的差别推断其作者归属。这其实和我们现在使用的垃圾邮件过滤器的原理是一样的。

这是一个典型的分类问题，如果要用机器学习来完成这个任务，我们就可以试试朴素贝叶斯。这里，可以假定对于一些常用词的词库，每个作者的习惯是不同的，这种偏好会体现在这些常用词的频率上。在最简单的情形下，作者就像抛撒硬币一样（只不过这一次的结果有 n 种），按照这种偏好随机地产生这样一些词。反过来，我们可以通过统计每个词出现的频率，来判断作者是谁的概率。

如何通过刚刚学过的最基础的贝叶斯决策框架完成这个任务？此处，我们有一个最简单的假设，就是特征与特征互相独立，每个词语的使用和其他词语没有关联。

而特征与分类目标存在可以测量的条件概率。我们制定一个决策的规则，这里假设整个文本生成的过程犹如一个伯努力实验，也就是以条件概率掷骰子，骰子的 6 个面就是常用词典。下一步，我们要模拟刚刚说的硬币游戏，根据已知数据，也就是在已知作者名字的文章里，利用词频统计，得到作者使用词汇的条件概率。如果有一个新的文章来了，我们就用这个从已知数据里求出来的条件概率表来计算该文章是不同作者的概率，最后选择那个概率较高的，认为作者就是他。

这里值得一提的是，整个模型的参数就是条件概率。而这里我们忽略了先验的存在，此处如果要加入先验也是完全可以的，就像我们在硬币游戏里做的事情，可以设定一个作者的使用偏好。

朴素贝叶斯还考虑基率的作用（你在人群里看到一个戴着眼镜的古板女生，你要判断她是商学院的还是数学院的，不仅要看她的特征，还要看商学院和数学院学生的人数，即预测要看数据本身的分布）。记录准确率，衡量模型的性能，如图 7.1 所示。

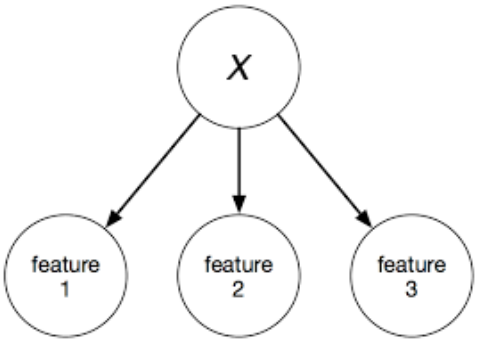


图 7.1

简单贝叶斯的本质就是一组特征离散且独立的分类问题。

简单贝叶斯的优点就是这个模型非常的简单，它特别适合我们在对问题的理解也特别简单的时候使用，另外一点就是它适合离散特征。所谓离散，就是我们选取的特征只有固定的几种状态，比如黑眼睛和蓝眼睛、某个单词是否在文章里出现。而身高体重是连续特征，这个模型就不那么适合了。模型里唯一可以变化的参数即条件概率本身，我们只需用刚刚讲的方法根据数据得到条件概率即可。

8 决策树方法

（阅读难度★★）

决策树处理信息的方法最像人，也是一个特别好理解的学习器。

人们在日常生活中处理问题的方法是按照一定的优先级处理特征，因为不同的特征总是以各种各样的方式相互关联。比如挑对象时，不会并列地处理好看、学历、有钱这些要素，而是根据心中的偏好按照一定顺序处理。例如，很多人是先看对象好看与否，才关注学历。决策树，就是用类似的思路解决问题的一个模型。

我们说决策树是我们接触的第一个非线性算法，非线性的第一个重要含义就是特征之间不是相互独立的，有人用 1+1 不等于 2 来形容。在分类决策的问题里，你需要考虑每一个特征对结果的影响，你无法按照一个固定的顺序，排列出这些特征哪一个对预测分类结果贡献最大。

那么，现在来看决策树是怎么简单地解决这个问题的。

决策树是一个非常接近人类思维的模型。形象地说就是一个调查问卷，把一个最终的决策转化为若干个问题，每步得到一个答案，按照答案的是和否来决定下一个问题是什么，如此形成一个树结构，最后形成一个分类器。

一般答案的回答是是和否，所以决策树体现了一种分而治之的思路，根据每个支点（树杈）决定后面树的走势。

比如你要买电脑，要根据很多特征来挑选电脑，比如 CPU、GPU、硬盘、内存等，你一定会问自己一系列问题，我要买哪款 CPU、GPU、硬盘、内存等，最后做出决策。你前面的问题一定会制约和影响后面的问题，并对后面问题产生影响。

CPU	GPU	内存	决策
高	中	低	买
高	高	中	买
高	中	低	不买

现在要做的就是把这个过程自动化，希望算法最后给出的答案是我们希望的判定结果。

之前的简单贝叶斯告诉我们如何把这些特征看成相互独立的条件，算出在不同条件下做决策的概率，而现在的过程是一步步来，每一步都是对某个属性的测试，每一步问询的最终结果不是给出最后的答案，就是给出进一步的问题，以帮我们减少不确定性，进一步考虑的限定范围一定是在上一步包含的可能性中。这个算法再次体现了机器学习模拟人的认知的特性。

这个过程很简单，但问题是，如何让机器在探索一个陌生的问题时有这样一个思维流程呢？我们希望机器可以学会对我们特定分类问题的思维流程，在一个陌生的数据到来的时候，也可以给出很好的结果。

如何把这个问题转化为机器学习问题，每个问题（比如你要什么样的 CPU、什么的 GPU、多大的内存和硬盘）都化作特征后，我们就得到了一个特征向量（CPU、GPU、内存、硬盘）。

数据是什么？数据是一张不同品牌、型号、特性的电脑的表格，最后我们的标签是电脑有没有被选中。

我们要构建一个决策树，让电脑可以自动做这个判断，如果人类要做这样一个选择，那么首先要做的第一件事就是对特征的关键性进行排序，顺序很重要。

如何选择这些次序？我们要想办法测量每个特征的重要性，或者说每个特征能给

我们提供多少信息。既然要测量信息，那么信息学的那一套工具就特别有用，这里不得不提一个被称作熵的概念。

什么是熵？熵就是对随机性的度量，测量的是我们随机变量的不确定程度，或者说我们对一个系统有多无知。比如，一个邋遢的男生比爱好整洁的女生的房间里的熵要高，因为男生房间里的东西处在高度无序和不确定的状态，需要很大的努力才能找到想要的东西。

如何使得熵减小呢？答案是整理房间。当你把无序的屋子变得有序，熵就减小了。

$$s = -\sum p_k \ln p_k$$

到这个阶段，我们需要一些数学知识，不要怕上面的公式，它说的无非是给定了一个随机变量的概率，熵就确定了。

它告诉我们，熵取决于两个条件，一个是问题包含了多少可能的结果，另一个是每个结果的概率。你的可能结果越多，每个可能的结果概率之间概率越均匀，你的熵越大。

为什么概率不相等熵更大？我们可以设想一件事情有两个可能，且可能性极端不平等：即一个是 0，一个是 1。大家都知道概率为 0 或 1 是一个确定的事件，没有随机性当然熵为 0，此时是熵最小的状态，反过来只有概率相等的时候熵才最大。为什么？因为我的信息量最少（就是刚刚说的，哪个可能都不比另一个更可能）。之所以有一个对数，是因为信息论里，如果你有 n 个可能且概率均等，对这些信息编码（给每个可能一个二进制的号码），需要的位数大体是 n 的对数，仅此而已。

回到决策树，我们希望率先找的特征，能最大程度地增大系统的确定性，使得系统的熵减少。我们用熵的变化来测量加入这个特征的效果，这就是我们建立决策树的原则。我们把熵的这个变化叫作信息增益，也就是增加的信息。

继续以电脑举例，假设一开始没有任何特征，那么对电脑买或不买的概率就等于买或不买的数量除以表格的数据总数。假设买与不买各占一半，此时你买电脑与否，该信息的熵为-(0.5×ln(0.5)+0.5×ln(0.5))= -1.38

下面加入 CPU 这个特征，对不同的 CPU 型号，你买或者不买的概率是 0.1 和 0.9，

这时候你就可以算出你的信息增益了。如果它比其他任何 GPU 等特征信息增益都大，那么第一个看的特征就是它了。依次在这个特征的每个可能取值下寻找第二个关键特征，列出第二个特征选的可能取值并寻找第三个特征，依此类推。对每一分支的操作里，如果我们发现在某个特征组合下的样本均为一类，则停止分叉的过程。整个操作过程形似寻找一棵不断分叉的树木，故名决策树。

决策树的最终状态，是希望每一个分叉下的样本都是一个类型，或者说纯度高，当然这个目标不可能完全达到，我们依然用条件熵对这样一个目标进行衡量，因此得到目标函数。

决策树具体要分支多少次？我们发现这个值的大小与问题的复杂度直接相关，由此可以看到机器学习问题里问题本身复杂性的重要性。

决策树最大的优点是可解释性很强。事实上，我们可以把它当作一个给特征重要性做排序筛选的方法，很多时候，我们只要看到树的前面几层，就可以对问题一目了然。比如，你做一个预测孩子身高的回归树，决策树的第一层可能是孩子的性别，决策树的左边是男生，右边是女生。决策树的分类过程酷似老专家和老医生的诊疗，因此如果你是一个机器学习工程师，要给客户解释一些东西，这是一个很好的方法。单棵决策树往往容易受到噪声干扰，由于它是一个个特征看下去的，人为的改变一些特征有时会造成判断错误，但是我们可以依靠构建集群模型来改变这点，这是一个机器学习非常实用的分支，把很多树以不同的形式连接，既可以构造随机森林，也可以构建梯度提升这样的串行方法，使得决策树的性能大大提升。

9 感知机：神经网络的基础

（阅读难度★★★）

感知机（Perceptron）是我们学习方法的第三大范式，也是仿生的最简单代表。

感知机的灵感取自神经元，一个神经元，由树突收集信息，轴突发射信息，好比一个最简单的决策机器，树突收集的正是那些刚刚说的要权衡的要素，而轴突就是根据输入是否超过一定阈值决定是放电还是不放电。整个生物神经元的树突和轴突都可以根据日常活动进行学习。

现如今，我们通常使用其他种类的人工神经元模型——感知机，感知机是一个最古老的神经元类型，直接采用阈值函数作为神经元的决策（激活）函数。当然还有各种其他类型的神经元，比如一种叫作 Sigmoid 神经元的神经元模型。那么，感知机是怎么工作的呢？简单来说，感知机的思维就是假设存在一条分类直线，然后这个神经元会测量错分点到这条线的距离之和，再用一个方法改变这条线的位置，直到把一类点保存到线的一侧为止。

感知机的输入是一系列连续的变量，在进行预测时，通过计算出的值是大于 0.5 还是小于 0.5（该值预期在 0~1）给出非黑即白的判断。

我们可以从几何上理解一下感知机是做什么的，它说的是求一个特征的线性组

合。还记得线性代数课上你学过的吗？如果我们把 w 放成 $w^2 = 1$ 的向量，这就是特征向量在一个方向上的投影！我们根据投影的大小对数据进行分类。从这一点来看，其实和刚刚讲的 LDA 差不多，差异在于背后假设和求解线的思路。

让我们看一个实际生活中的例子，智能助手推荐你去看某一部电影，这背后的逻辑就是你要一个机器帮你做决策去不去看电影。当然一开始它不了解你，但是它可以先帮你做决策，然后根据你每次给它的反馈意见学到你的偏好，这就是感知机训练的方法。如果你给它肯定意见，它就不做改变；你不满意，它就变一变。

如果你来设计这样一个算法，要怎么做呢？首先，写出目标函数，什么是合适的目标函数？就是让训练的结果是你想要的。在这里，你要让你的预测少犯错，也就是错误最小。什么是错误最小？就是分类错误的数据在 w 方向的投影值最小。如果是 0 的话，就是没有错误。如果得不到 0，也要它最小。

假设你让机器决定是否去公园的信息包括天下不下雨，女朋友在不在家，然后某一次，天下雨，但它让你去公园，结果你发誓下次机器再这么笨，你就拔电源了，于是机器就会减少这个要素的权重，这样一来，它下次让你去公园的概率就会下降，直到你满意为止。就是这么简单。上述的更新和优化的方法，被称为 SGD（随机梯度下降）。

感知机的一个大缺点是太绝对了，除了是就是否，除了黑就是白。

在真实情况下，这一定是不对的。首先，真实情况从来就不是非黑即白，而是存在灰色地带。比如今天天气不好，但是你女朋友恰好今天有空，你决定到底去不去公园，可能真的要投一个硬币来决定了。假设数据集不能被一条线很好地分开，而是在这条线周围有很多重合的点，这时候你硬要用一条线来分割，而有些时候，就不太好玩了，比如判定疾病。在某些条件下，得不得病真是不好判断，如果你硬要给人判定是恶性肿瘤或是良性肿瘤，这可能是人命关天。比较负责的方法是告诉人家恶性或良性的概率，具体采不采取手术，患者自己斟酌。

这时候，我们需要一个算法，不仅实现分类，还要告诉我们概率是多大。这时候，我们有一个强大的方法，叫 Logistic 回归。在上文叙述的感知机中，用一个阈值函数来判定分类，如同一个拿着刀的裁判，而这一次，我们要把这把刀替换成渐变的分类

界面，引入一个平滑曲线替代之前的阈值，如图 9.1 所示。这就是 Sigmoid 函数，也是最为常用的求分类概率的方法。

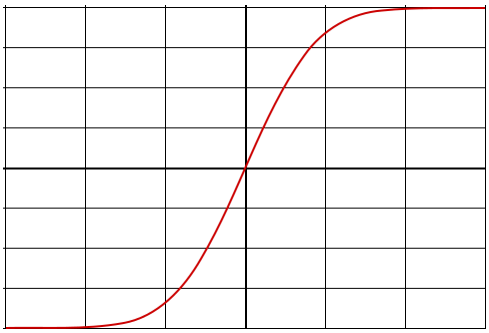


图 9.1

下面看看如何求得 Sigmoid，直觉理解是把这个分类界面柔化，用概率来表述界面附近分属于两类的概率，从一类到二类一定存在某个平滑过渡。关键是我们采用什么样的函数柔滑？我们知道高斯函数还是比较特殊的，是否存在一类更普通的函数让我们拟合？

这个函数正是我们之前说的 Sigmoid。Logistic 回归的本质思维是，用一个最通用的拟合分布函数，来逼近任意一个未知分布。我们寻找一个能够衡量数据和概率模型匹配度的函数来看 Logistic 回归的拟合度，这个工具就是似然函数。假定那些数据实际都是这个概率分布，也就是数据的分布，然后看有多大概率观测到这个数据，概率越大，说明 Logistic 回归越符合真实，于是，就有了优化对象。

刚刚讲的一个神经元是个线性分类器，那么如果问题改成在特征平面上画一个拐弯的线，或者画个甜甜圈呢？

这时仿生主义模型将大放光彩，我们将得到由很多感知机组成的神经网络，它的发展直接指向深度学习。

10 降维：应对复杂的通用武器

（阅读难度★）

科学的进步，商业的发展，得益于数据所统领的疆域的扩展。

大数据不单单是指数据量大，也不仅仅是指数据来源广（数据的维数大），还意味着数据间的关系更加复杂（复杂网络的涌现）。数据中既包含有用的信号，也包含无用的噪音。但不同维度的数据可以互相验证，互为因果，从而带来数据间的有序性。数据在不断变化中，而变化的趋势又受制于数据之外的环境因素，这些环境因素具有自我指称的特性。

数据的扩大并不会必然带来视野和格局的扩大。而数据来源一旦多元，就一定要有去伪存真的步骤，去除重复，剔除噪音，清洗缺失值，使数据同步、准确、完整。这是当前数据处理中最耗时的步骤，而如何避免在数据清洗时过度补偿，如何避免丢弃过多有用的数据，则是需要智慧的。

面对庞杂的数据，降低维度似乎是一个习惯性的处理模式。降维是使用较少的变量来概括性描述目标数据的特征，而降维的极致是将种种复杂的情况变为一个数。例

如，对于得了癌症的患者，他们最想知道的是平均还能活几年。但是，这个终极的降维将会扭曲过多的信息。

这个问题的实质，是问问题的人不想也不会知道全部的信息。资料库中这类患者共有数百万人，每个人接受的治疗方式不同，每个人本身的身体素质不同，每个人的心态也不同，从而导致每个人诊断后的寿命也不同。但问问题的人不是在做科学研究，他只想基于短小明确的信息做出判断。这个时候，有用的信息必须要在一分钟之内说完。

要想涵盖更多的信息，我们首先要做的就是分类。而对于一个随机变量，最关键的不是它的统计指标，而是它的分布。比较有名的分布有正态分布和指数分布两种，不同的分布，其结果天差地别。正态分布对称，尾巴不长，不太可能出现过于离谱的事件。而对于指数分布，远离平均值的事件注定会发生。

对于癌症患者来说，搞清楚存活时间的分布远比搞清楚平均存活时间重要得多。如果存活时间是正态分布的，那么有 50%的概率活过平均存活时间，之后死神会越来越近。而如果是指数分布，由于其 scale free（无标度）的性质，当活过了平均存活时间后，预期还能再活一个平均寿命那么长，也就是说，死神会越来越远。正如《反脆弱》中所说，如果一个事情存活了 1000 年，那么你可以预期它还可以再存活 1000 年。

事实上，癌症患者的存活时间更接近指数分布。理解这则信息和其背后的意义，以及对于癌症患者的意义，远远超过了平均存活时间。这则信息说明，同样的治疗方法，相似的病症，却有着相差甚大的存活时间。其中肯定有些正反馈存在，不然不会有这样一个尺度一致的分布。而这个正反馈，最有可能的便是患者的心态和求生意志。更强的求生意志会带来更多的存活时间，而这会延长预期存活时间。

从癌症这个例子，我们可以看到降维的问题之一是忽略分布，而另一个问题则是选择的指标可能不够基础。例如，按照流行的观点，对于学生来说，一维的指标可能意味着分数至上；而对于企业来说，则可能是利益至上。

有的人不同意这样的降维方法，说评价学生指标的应该是综合素质，20%的体育，20%的艺术，30%的修辞，30%的科学，也有人认为这个比例不对。

降维的目标是为了给出一个排名，而排名是为了给出是否进步的评价，然而，自然界的进化并非如此。

不同的基因是不同的维度，每一种基因的组合都会带来不同的适应性，进化的目的是找出最优的基因组合，使生物体有最优的适应性，不是跑得快的就一定能抓到猎物。

生物体的降低维度，不是对几个维度的加减乘除，而是通过一套复杂的大网，将个体的性状变成基因组合的适应度，从而指导生物的进化。

比如说中国文化对世界文化的贡献，可选取的维度有中文书报的数量、中文论文的数量、使用中文的人数、中文网页的数量等。要想降低维度，我们需要的是让各国的文献互动，最后选出的维度是有多少中文被翻译成了外文，然后再来评价各国文字相互翻译的表格上中文出现的中心度。

PageRank 算法也是在降低维度，从而形成对网站的排名。总统大选也是在降维，最终落到了从两个人中选择。

PCA 之所以能降低维度，是找到了 A、B 两个数据点的相关冗余的部分，如果待降维的点都在坐标轴上，那么降维损失的信息量就很大了。之所以这样，是因为真实世界的问题往往看似高维（复杂），但由于自然或者社会本身的各种约束，真实的维度并没有那么高。

这里的原因在于特征很多时候并不独立，而是有一定的限制关系。比如 *x1* 和 *x2* 组成的特征平面，两个特征互相限制，此消彼长，实际的方程是 $x1^2+x2^2=1$，所以真实的维度其实是一维。

比如说房价问题，看似房价取决于房屋面积、各类配套设施、房屋地点环境、地点富裕度、地点治安，等等。但事实上很多要素其实都是相关的，比如卧室数量和配套设施、房屋地点环境和富裕度等。如果有一个有效的方法把它综合成房屋本身大小指数、地点指数、配套设施指数，则很可能使得问题大大简化，而且几乎完全保留了问题的真实。

再举个图像处理的例子，如果把人脸图像的每个像素看成一个特征，那么其整体

会组成非常大的向量，仅维度就有一千万（手机的像素多在一千万以上），那么每个维度即使只取 0 和 1，那么所能得到的可能性也是天文数字。但是相对我们的识别问题，图像里的内容的种类并没有那么多，远远低于这些特征所具有的可能性。正是因为并不是每一种特征组合都是合法的存在，所以大部分时候，特征之间的关联使得可能的组合数大大降低，实际存在的自由维度比想象的要小很多。你可以想象在一张图片上做平移、压缩、旋转等几种变换得到一系列新的图像，但是它们表现的内容不会有太大变化，人还是那个人。我们实际是在高维空间一个极小的紧闭区域里变化。大家可以想象一下刚刚列举的那个二维空间里，由于约束导致我们最终得到的特征空间是一个一维圆周的问题。

对于这类问题，我们可以使用降维算法得出一个较低维度的问题，从而使所需要训练的数据大大减少。这条路指向了无监督学习，比如 LDA、PCA 和各类 Embedding 的问题。

为了达到这个点，我们需要对特征不停地进行编码，有时候叫作表征，有点像某种形式的坐标转换，最终在新的空间里让问题变得更简单。

但是这么做的代价是我们通常需要根据数据的分布猜测一个函数结构，比如线性结构，或者我们熟悉的多项式拟合及傅里叶变化，然后参数可以由数据反求出来，如果猜测是正确的，就可以用很少的数据预测非常多其他的数据。

在生活中，要想减少过度降维的弊端，就需要找出相互之间独立的维度，然后通过现实生活中的竞争，动态调整不同维度间的权衡。

对于物种，这两个独立的维度是吸引异性的魅力和让自己活下去的能力。

对于企业，是花在探索新领域（exploration）和花在巩固已有疆土的能力（exploitation）。

对于学生，是之后继续学习的能力和已熟练掌握的知识。

多找出几个相互独立的维度，才能避免被过度降维成一个数字。

不要总用一个简单的指标来评价别人，久而久之，对自己的评价也只使用那个孤立的指标。

所有的降维，都会损失一部分原始信息。就像可以用中位数、众数、平均数来为一组随机变量降维，但无论使用哪种降维方法，都有各自的问题。只有明白了随机变量的分布是更趋近正态分布还是指数分布，才有可能减少降维带来的信息损失。

代替降维的最优方法是先构建一个相互作用的网络，看看网络中哪个节点处在中心。节点的互动所包含的隐藏信息可以在网络的构建中被展示出来，从而使降维所依赖的信息不只是变量之间的相关性。不过这需要来源多样、时间连续的数据，而这样的数据又受环境因素、采集过程等因素影响，从而影响降维的结果。

至此我们已经介绍了几种非常简单的模型，至于复杂的核心思想则是降维。在纷繁里寻找最关键的因子，奉行的原则是如无必要，勿增实体。后面，我们再来看看用复杂对抗复杂的方法。既然自然演化出来的对抗复杂的最有力工具是人类大脑，那么人类的机器学习工具又何尝不试试更复杂的呢？

第三部分　神经网络

11 神经网络不神秘

神经网络为什么强大（阅读难度★★）

世界是非线性的，机器能否阐述复杂问题，关键在于能否处理这类问题。

举个很多人都关心的例子吧——机器学习能否预测房价。我们把问题简化为卧室数量、面积、地段，然后求成交价。这里用线性模型求解价格，假定卧室数量、面积、地段都会独立地影响房价，所以先考虑哪个特征都可以。而真实的情况可能是地段只对大户型和小户型的房子有影响，毕竟不同户型的房子针对的用户群是不同的，比如买大户型的人群和买小户型的人群对地段的偏好就不太一样。

如何处理这种复杂的信息呢？你可能会想到让不同的区域，享受不同的权重。好比我们有好几位房价评估师，每一位都是针对某种情况下的房价的专家，再把这些专家的想法综合起来，如图 11.1 所示。

那么神经网络是怎么完成这个工作的呢，将不同的意见合并起来了？我们需要做的是利用一个非线性函数实现筛选，这就是我们需要填入的第二个要素——激活函数，也就是图 11.1 右侧单个圈里的内容。它所做的事情，就是根据一定条件决定前面过来的信息放不放行，也就是对交来的信息（房子的特征）进行分流，这有点像之前是金子和沙子全都混在一起，而现在要把金子和沙子分开。我们可以假设每一个绿

色的圈（中间层）都是阈值线性函数，而不是直接线性加和，这样一个小小的非线性，使得某一组权重只在某个条件下才能启动（超过阈值），这样，我们就实现了一个根据不同区域条件进行加权考虑的模型，从而得到最终的价格。

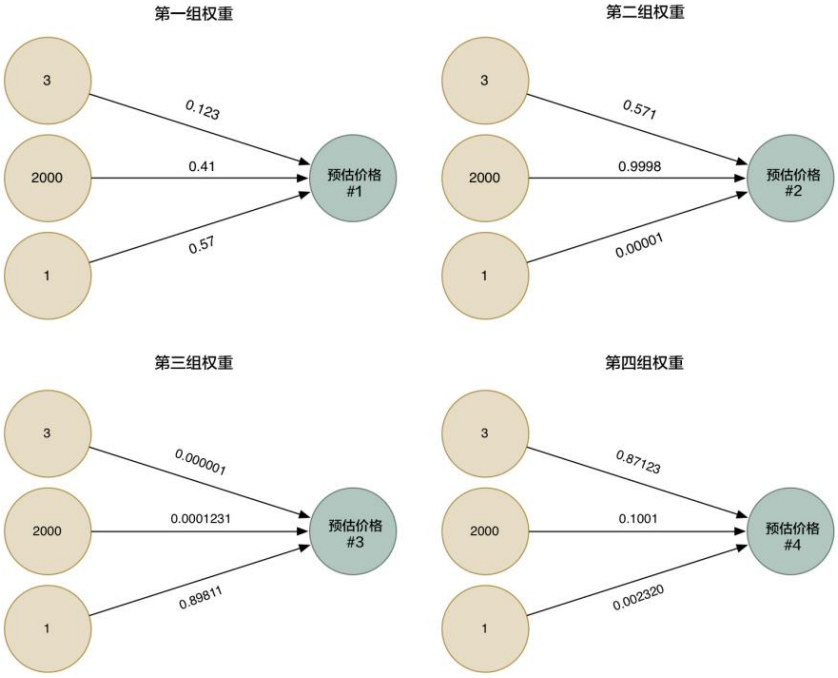


图 11.1

我们知道感知机可以对特定的特征组合进行编码，但事实上，感知机始终解决不了一个问题，就是或非门。一个或非门可以接收两个输入信号，若是这两个信号是相同的，则或非门输出 0；如果不同，则或非门输出 1。这是一个入门级别的非线性分类，下面看看如何通过最小的改动，来实现这个功能。

这个问题事实上十分常见！假设我们所说的或非门是一个三明治结构，之前我们切一刀或用一条线就可以把平面分成两部分，而现在用两条平行的直线就可以按照或非门的要求把平面分成三部分了。我们还是有两个输入特征，然后一左一右配备两个感知机：一个是权重（1,1），另一个是（-1,-1）。这样便拥有了一个完美的三明治结构！神经网络，一个是多层，另一个是非线性的 gating，将这两个组合起来就可以

解决各种复杂的问题。

似乎把简单的感知机叠加起来，就可以解决许多复杂的问题。但事实上，神经网络已日益成为通用的机器学习工具，为什么有这样的能力呢？

神经网络其实是三个元素的叠加：

（1）线性变换；

（2）非线性变换；

（3）线性决策。

这三点反复叠加就可以了。

（1）和（2）我们事实上实现的是一个坐标变换，让我们把信息聚合到一个更加好分类的空间，然后按照之前的非线性函数进行分类。

换句话说，就是特征变换（降维），具体来说就是表征学习（Representative Learning）。表征学习顾名思义，就是把一种特征转化为另外一种特征，类似于 PCA 所做的事情，只不过是非线性的。从一个角度理解多级神经网络所做的事情就是表征学习。这也是人们常说的深度学习可以把机器学习里最麻烦的特征工程自动化。

下面再仔细理解一下这个特征变换的过程。

想象一下，线性变换可以做的操作是坐标旋转、降维或升维，以及平移，然后输入非线性函数进行扭曲、折叠等操作，可以用上面的例子来理解一下。

我们实际做的是一个坐标变换，使得整个数据集有了一个新的表征，在这个表征之上，之前线性不可分的东西变得线性可分了。

如果要分类一些更复杂的边界呢，如同 SVM 能做的事情？没问题，这里我们用到的一个思维就是集群思维。既然说一个神经元是一个分类器，那么这里把很多个神经元组合在一起，就可以得到很多个分类界面。在此之上，我们再引入一个决策层，决策层具有的功能是取交集，就可以得到被众多分类界限包裹的问题。事实上我们可以证明，一个一隐层的神经网络，可以模拟任意复杂的连续边界。

那么，是不是只有阶段函数的隐层激活函数才可以做到呢？不是的，事实上，任

意的非线性函数都可以做到。为什么？我们来看看神经网络三要素里非线性的重大意义。这里的问题某种意义上也是升维打击，原先有 *x1* 和 *x2*，这里可以给它通过线性变换加上非线性，得到一个高维度的空间，在此之上，我们可以用一个超平面来做任意地分割，与支持向量机的思路类似。这里的核心还是非线性，假如我是线性的，则将会导致一个灾难性的结果，也就是说，我做再多的操作，其实还是二维的，你只需写一下你的总变换函数即可。

下面谈谈激活函数的选择，在最简单的神经网络里，激活函数就是感知机用的那个 Heaviside（赫维赛德阶跃函数），但是，这个函数是不可导的，它将导致根本无法训练。怎么办呢？还记得 Logistic 回归吗，引入光滑的 Sigmoid 函数，就能解决这个问题。Sigmoid 函数在所有点光滑连续可导，符合我们的要求。然而为了方便，我们更偏爱那些 mean 为 0 的向量，所以我们在实际使用的隐藏激活函数里更常用的是它的改进版，也就是 Tanh 函数，这个函数以零为中心，方便操作。

而事实上，这依然不是我们目前最常用的激活函数，更为常见的激活函数是：ReLU。其实各个非线性是等价的，但是，神经网络一定要有的思路是：梯度是要在层之间传递的。我希望我的梯度信号在各层里传递的损失越小越好。什么是最保真的变换？当然是线性，也就是它本身，但是我还要保留一个非线性，因此，我选择一个非线性函数里最接近线性的函数，也就是 ReLU，max（0,z），正的部分保真，负的部分去掉，这样既能起到非线性的作用，又可以做到保真。

如果按照现在的角度讲下去，那么只需要一个超大的浅层网络即可，从中实现一个从低维到超高维度的映射，然后在这个超高维度的空间表征做任意一件事，而这正是最近几年神经网络横扫四方的原因所在。你可以用任意多的神经元逼近任何复杂关系，但是，这里面有如下两个问题。

首先，大家还记得维度灾难吗？在你给出这个解决方法的时候，新的表征维度变得无穷高，使得你需要足够大的数据量来解决问题，这是机器学习的终极难题。如果要编码一个 *n* 维的特征数据，那么你通常要做的是用 exp(*n*)的神经元来 shatter the data（2××2××n）（粉碎数据），所以你需要 2 的 *N* 次幂的神经元去表示所有可能的模式。这样的效率可想而知。

其次，还有一个致命的问题：那就是过拟合！这样靠一条条线拼出来的图像，靠一一对每个数据局部进行拟合的结果是，算法只会死记硬背，无法理解任何其中的结构，也就无法找到预测性。

有什么方法可以克服这个问题？引入深度。用一个比较好懂的方法讲，更多的层数代表更高效率，因为它能够充分利用隐层的特征变换功能，像 $L(g(f(x)))$，这样的层级迭代，我们有时候叫作层层抽象。可以从数学上来理解这件事，同样是 10 个神经元，你把它分为 5+5 和 10，本身能够编码的信息是不同的。

认知科学里说，抽象使得不同的事物可以在更深的层次上被联系起来，也就是说，不同层次的函数迭代，可以产生新的空间（某种程度也是降维，比如图形处理中），使得原先的距离概念被打破。具体怎样被打破，这非常像人脑处理信息。实验表明，越靠近感官的神经元处理的信息越具体，比如颜色、亮度，而经过多级处理，到达皮层之后，我们的神经元就可以对比较抽象的实体，如人名、物体的种类敏感。

多层神经网络的训练是困难的，矩阵里的元素权重很多，而这个时候，对整个网络的训练就显得难度比较大，我们常用的方法是梯度回传。这个时候的梯度依然是我们之前感知机讲的梯度，但有所不同的是，这个时候的梯度需要在多层网络里传递，从信息的角度来看，就是要把错误的信号传递到任何影响预测的参数里。通常我们把这个信息回传的过程写成一个连续乘的过程，这个过程里，我们的梯度算法可以对所有参数进行更新。通常神经网络是输入—线性组合—非线性的三明治结构的不停迭代，因此，我们的回传是这个三三结构反过来相乘，从输出到非线性，从非线性到线性，每次我们都根据梯度对矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 进行更新。在这个过程中，需要注意的是，神经网络的求解往往不是一个，而是多个最小值，我们通常最后要的不是选出一个最小值，而是找到一个合适的极小值。

适当的目标函数加上所有的正则化方法，就构成了深度学习的基础。

最后讲一个神经网络中的常见异型：Auto-Encoder 自编码器，这是一个把自己映射到它自己的模型。一个 Auto-Encoder 把输入变换成自己的另一种形式，我们通常让它做的事情就是降维，如同 PCA 一样，在降维的过程里优化一个 Loss Function，$L(x,g(f(x)))$，就是把 x 翻译过去，再翻译回来，让这个语义的损失最小。

而我们这里与 PCA 不同的是，我们所做的是一个非线性的降维处理，而不仅仅是矩阵做的空间变换。所谓编码，说的是我可以用一个代码来替换这个数据，但是编码里并不一定含有元数据的结构信息，比如你给每个图片从 1 排到末尾，但是这些数字并不包含任何有关源编码的信息。在表征学习里，我们希望得到一个低维度且保持原数据结构的表示。这样做有什么用处，且到 RNN 的部分再细细说来。

12 CNN 的几个关键词

（阅读难度★★★）

你不得不知道的 CNN

说起 CNN（卷积神经网络，Convolutional Neural Network），最初人们想到的都是某电视台，也许几年之后，人们想起的多半是深度学习了。

应该说，CNN 引起了近几年深度学习的风暴。2012 年，正是它让打入冷宫的神经网络重见天日，并且建立起自己在人工智能王国的霸主地位。

如果你认为深度学习只能用来理解图像，那就大错特错了，因为它的用途太广了，包括文字、图像，还有音频，从手写数字识别到大名鼎鼎的 GAN 对抗学习，都离不开它。

不过要想了解 CNN，还是拿图像做例子比较恰当。用一句话来总结 CNN 图像处理的本质，就是信息抽取，巨大的网络可以通过抽取，一步步得到最关键的图像特征，我们有时也称之为自动的特征工程。

CNN 的建造灵感来源于人类对视觉信息的识别过程，如图 12.1 所示。人脑对物体的识别的第一个问题是：对应某一类对象的图像成千上万，比如一个苹果，就有各

种各样的状态，我们识别物体的类别，事实上是给这成千上万张不同的图片都打上同一个标签。

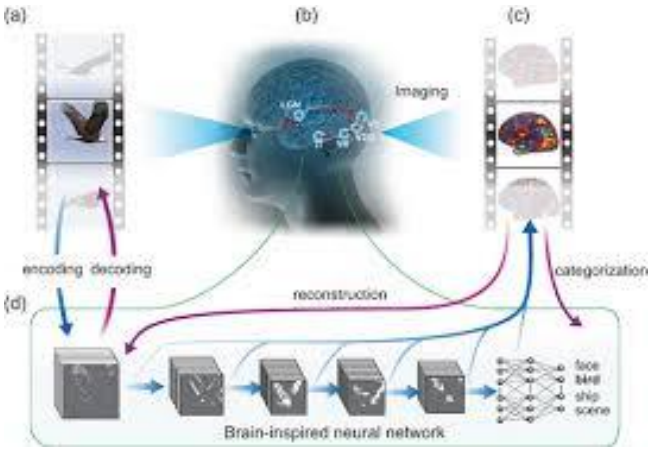


图 12.1

物理里把这种一个事物的结果与一系列的变化都没有关的特性，叫不变性。比如，如果你转动一个苹果，任何一个角度它都是苹果，这就是苹果有旋转不变性；但是数字 6 就不行，如果你给它旋转 180 度它就变成 9 了，它就没有旋转不变性。

我们通常可以无视这些变化认出事物来，也就是把和这种变化有关的信息忽略。如果我们对图像进行识别，事实上我们的算法就要有人的这种本领，首先让它学会什么东西与真实的物体信息是无关的。

就拿数字识别举例，一个数字是什么，虽然与旋转的角度有关，但是与它在图片中的上下左右位置无关，我们称这种不变性为平移不变性。

要想解决这个问题，最粗暴的一个方法就是制造很多的样本，比如把“1”放在很多不同的位置，然后让机器在错误中学习，直到穷尽所有的位置。不过几乎没有人是这样完成对物体的识别的。

那应该怎么办呢？CNN 中的卷积正是这一问题的答案，因为卷积操作本身具有平移不变性。

卷积，顾名思义，“卷”有席卷的意思，“积”有乘积的意思。卷积实质上是用

一个叫 Kernel 的矩阵，从图像的小块上一一贴过去，一次和图像块的每一个像素乘积得到一个 output 值，扫过之后就得到一个新的图像。下面用一个 3×3 的卷积卷过一个 4×4 的图像，看看取得的效果，如图 12.2 所示。

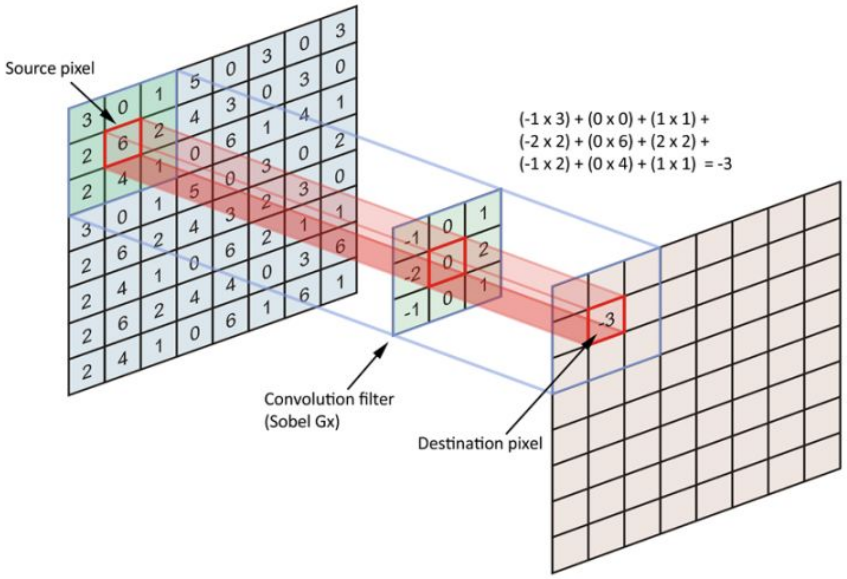


图 12.2

卷积的数学过程

一个卷积核就像一个小小的探测器，它的 DNA 是被刻录在卷积核的数字里的，告诉我们它要做什么，而卷积核扫过图片，只要它的 DNA 是不变的，那么它在图片上下左右任意哪个位置看到的结果都是相同的，这便是卷积本身具有平移不变性的原理。由于这种不变性，一个能够识别 1 的卷积在图片的哪个位置都可以识别 1，一次训练成本，即可对任何图片进行操作。

在图像处理领域，卷积早已有另一个名字，叫作滤镜。我们把图像放进去，它就出来一个新图像，可以是图像的边缘，也可以是锐化过的图像，还可以是模糊过的图像。

在 Photoshop 里面有一些滤镜，比如说锐化、模糊、高反差识别这一类，都是用类似的技术。这类技术所做的事情是，对图像的每个小片用一个矩阵进行处理，得到

一个画面的转换。我们有时候会说低通和高通滤镜，低通滤镜通常可以用来降噪，而高通滤镜则可以得到图像的细微纹理。Photoshop 应用的就是卷积，卷积核里面的数字定了，它的功能也就定了。

为什么这样做有效果了？因为图像的特征往往存在于相邻像素之间，Kernel 就是通过计算小区域内像素的关系来提取局部特征的，可以理解为一个局部信息的传感器，或物理里的算子。

比如边缘提取滤镜，它所做的物理操作又称为拉普拉斯，像素只有在从明亮到变暗的过程中它才得 1，其他均得 0，因此它所提取的图像特征就是边缘。事实上，我们知道，图像中的信息往往包含在其边缘，你给一个人画素描，只通过大致轮廓一定能够完全识别这个人。我们通过寻找得到信息的关键载体——边缘，而把其他多余的信息过滤掉，得到了比第一层更好处理的图像，大大减少了需要搜索图像的可能性。

卷积的边缘抽取过程如图 12.3 所示。

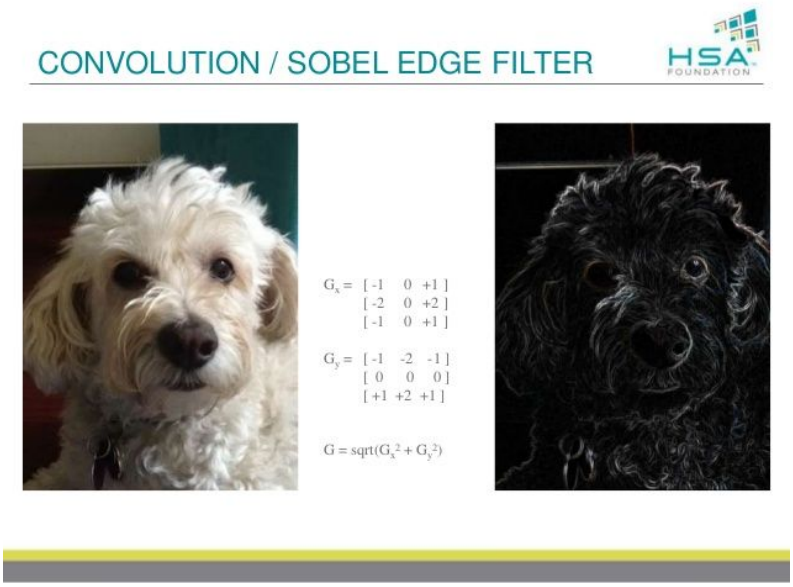


图 12.3

常用于卷积的 Kernel 本质是两个：第一，Kernel 具有局域性，即只对图像中的局部区域敏感；第二，权重共享。也就是说，我们是用一个 Kernel 来扫描整个图像

的，在此过程中，Kernel 的值是不变的。这点就可以保证刚刚说的平移不变性。比如说识别一个物体，显然这里的识别不应该依赖于物体的位置，而是和位置无关，即平移不变。

那卷积如何从不同的图形中识别数字 1 呢？数字的尖锐的线条会让卷积的值很高（响起警报）。无论 1 出现在图像中的哪个位置，用局部扫描+统一权重算法都可以找出来，用同一个识别 1 的卷积核来扫过图片，任何一个位置的它都可以找出来。

卷积和神经网络有什么关系呢？答案是卷积扫过图像时，每一个卷积核与图像块相乘的过程，都可以看作一个独立的神经元用它的神经突触去探测图像的一个小局部，然后再做一个决策，就是我看到了什么或没看到什么。整个卷积过程相当于一层神经网络。

刚刚提过，卷积是一个能够对图片中任何位置的同一类信息进行抽取的工具，还讲到除了抽取，另外要做的一个工作是取出重要的信息，扔掉不重要的信息，实现这个操作的，叫作 Pooling。

但是请注意，这个时候如果原图像的大小是 28×28，那么从 Kernel 里出来的图像大小依然是 28×28，事实上，大部分时候一个图像的局部特征的变化都不会是像素级。我们可以把局部特征不变性看作一个假设，把这个假设作为一个数学公式加到卷积层里，帮助我们过滤冗余信息，这就是 Pooling 所做的事情，也就是扔掉你周边和你长得差不多的那些像素。

Pooling 的本质是采样，以提升统计效率，即利用局部特征不变性降维。Pooling 的方法有很多，常见的叫作 Max Pooling，就是找到相邻几个像素里值最大的那个作为代表，把其他的扔掉，如图 12.4 所示。

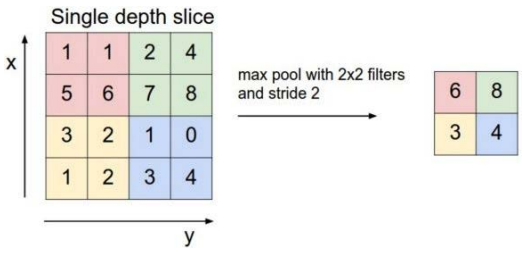


图 12.4

这样经过从卷积到 Pooling 的过程后，在识别 1 的任务里，我们可以验明在每个小区域里是否存在边缘，从而找到可能存在 1 的区域。在 Pooling 的终结点，我们得到的是一个降低维度了的图像，这个图像的含义是告诉你在原有的图像的每个区域里是否含有 1，又叫作特征图。

好了，我们可以从一堆图片中识别出 1 了，那么怎么识别 2 呢？我们把 2 写成一个 Z 型，应如何做到这点呢？我们不仅要学会识别竖着的线条，还需要学会识别横向的线条，记住，一个卷积层只识别一个特征。如果既要找竖线，又要找横线，那么就需要两个不同的卷积层，并且把它们并联在一起。

接下来，横线对应一张特征图，竖线对应另一张特征图，如果要识别 2，就需要比较这两张特征图，看是否有哪个位置两个特征图同时发生了警报（既有横线又有竖线）。

这个比较的过程，也可以用一个卷积搞定（依然是平移不变性）。

这时候，新的卷积层对之前并联的两个卷积的结果做了一个综合，或者说形成了一个特征的特征，即横线和竖线交叉的特征。手写数字识别如图 12.5 所示。

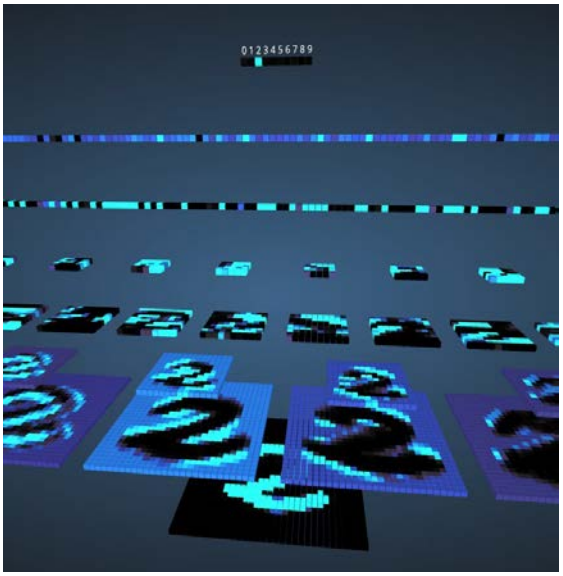


图 12.5

这里把我们的理论更上一层楼。深度意味着什么？要想正确识别一个图像，既不可能只看边，也不可能只看角，我们需要对图像的整体有认知，才能知道是张三还是李四。也就是说，我们要从局部关联进化到全局关联，真实的图像一定是有一个全局的，比如我的脸，只有我的眼睛、鼻子、耳朵都被一起观察时才称得上是我的脸。

从一个层次到另一个层次的递进，通常是对上一层次做横向及纵向的整合（图层间的组合或图层之内的组合或两者都有）。我们的特征组合是基于之前已经通过 Pooling 降低维度的图层，因此事实上每一个神经元决策的信息相对上一层都更多。我们用一个学术名词——感受野，来表述一个神经元决策所涵盖的像素多少，上一层次看到更多的输入神经元，因此感受野看更多了。越靠近顶层的神经元，所要做的事情就越接近全局关联。

感受野越大，表示越抽象，如图 12.6 所示。

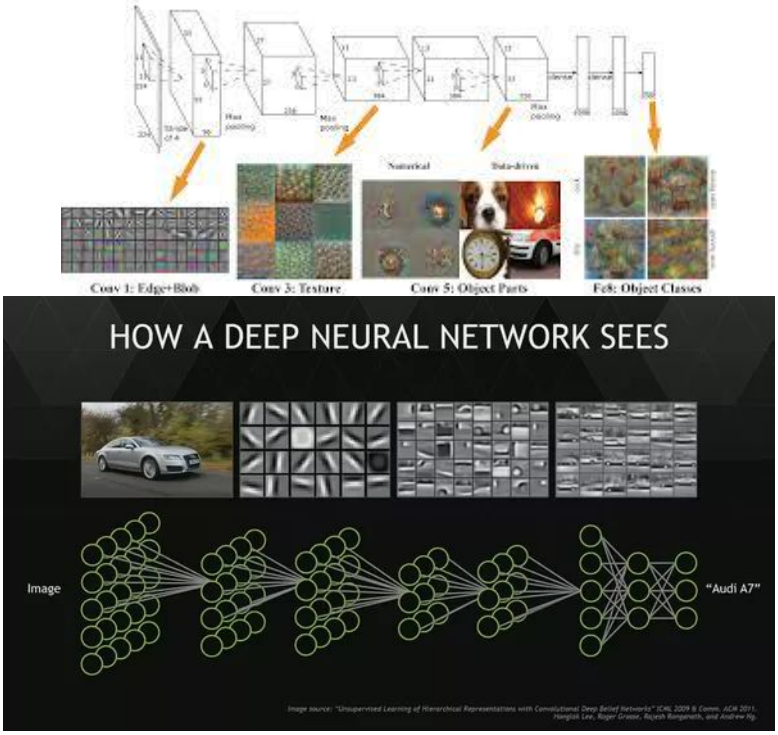


图 12.6

这和物理学的一个基本方法——尺度变换有着异曲同工之妙（本书后面讲），也是提取全局信息的一个非常核心的方法，这里称之为层级递进法。一级一级的对画面进行降采样，把图像里的四个小格子合成一个，再把新的图像里的四个小格子合成一个，直到一个很大的图像被缩小成一个小样。每一层的卷积，都不是一个卷积，而是一组执行不同特征提取的卷积网络。比如刚刚说的，不同方向的边缘构成的一组卷积，你可以想象后面有不同大小的角度组成的一组网络，它体现了在一个空间上我们所能够达到的特征工程。

如此级级互联，越靠近上层感受野就越大。整个 CNN 网络如同一个封建等级社会，最上层的是君王，它是唯一具有全局视野的人，下一级别，是各大领主，然后是骑士，直到农民（底层神经元）。

我们把刚刚讲的全局换一个词，改为抽象。深度卷积赋予了神经网络以抽象能力。这样的一级级向上卷积做基变换的过程，和人的思维做个比喻就是抽象。抽象是在很深的层次把不同的东西联系起来，CNN 教会了我们事先抽象的一种物理方法。

到目前为止，我们所描述的都是一些人工的特征工程，即使网络再深，也顶多是深度网络，与学习无关。这样一个系统，如果要人工设计，就算我们穷经皓首也可能做的都是错的。因此，这样的一个结构，只能靠机器自己学，这就是深度学习的本质了。我们通过几条基本假设（正则）和一个优化函数，让优化（进化）来寻找这样一个结构。基本假设无非是图像的几个基本结构，体现在几个不变性上，物理真的是很伟大啊。

深度学习的训练，就是计算机帮助人完成了机器学习最难的一步特征工程（特征工程本质就是基变换）。以前人类绞尽脑汁思考如何做图像识别，寻找人是如何识别图像的，希望把人用来识别物体的特征输入给计算机，但是现在通过深度卷积，计算机自己完成了这个过程。

卷积网络大家可以关注以下几个方向。

（1）更深的模型。从 AlexNet 到 VGG 19、High Way Network 再到残差网络，一个主要的发展趋势是更深的模型。当你采用更深的模型时，通常可以发现一些神奇的事情发生了。当然，网络的宽度（通道数量）也在增加。

VGG 19 的结构如图 12.7 所示。

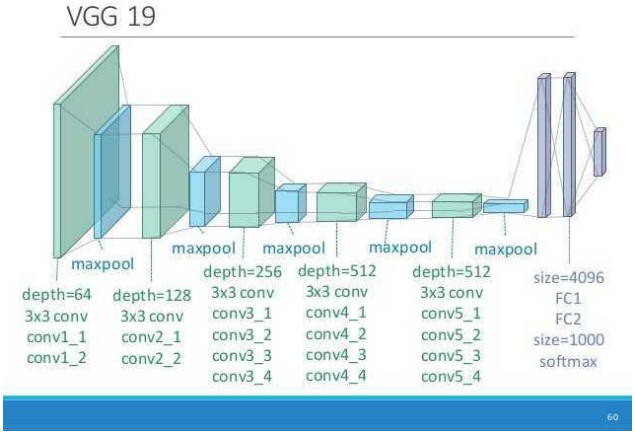


图 12.7

GoogLeNet 的结构如图 12.8 所示。

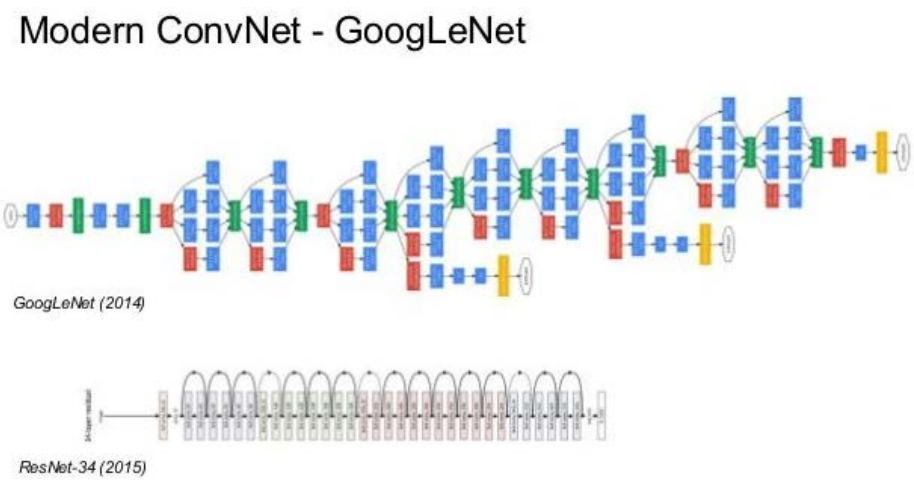


图 12.8

（2）更通畅的信息交换。深，带来的第一个问题是训练困难，反向传播难以传递。从残差网络，到 Dense Network（如图 12.9 所示），一个主要的发展趋势是不同层级间信息的交换越来越通畅。我们逐步在不同层之间加入信息的直连通道。

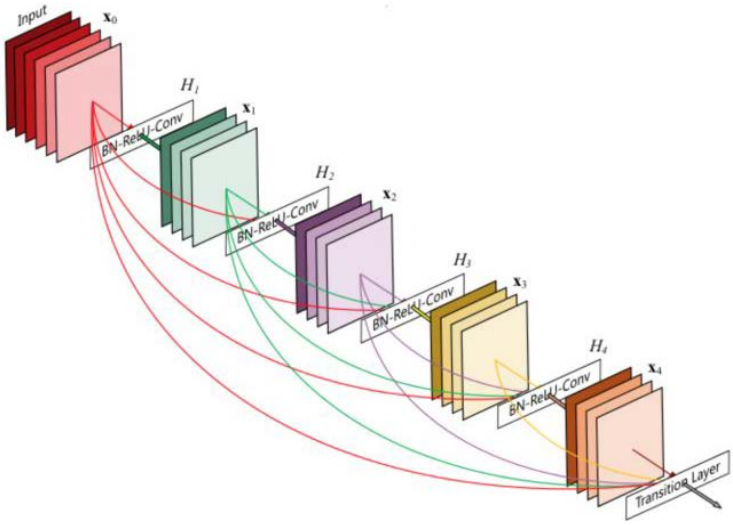


图 12.9

（3）与监督学习之外的学习方法的结合，如迁移学习、半监督学习、对抗学习和强化学习。后两者的有趣程度远超监督学习。

（4）轻量化。CNN 网络越来越深，使得网络的文件动辄装不下，这点使得 CNN 网络的轻量化部署成为重点，我们希望在性能和能耗间折中。一个很好的方法是对网络权重进行减枝，去掉不重要的权重，如图 12.10 所示。另外一个方法是缩减每个权重的数据位数，甚至只用 0 和 1 表示。虽然看上去我们丢失了很多信息，但是由于巨大网络中的信息是统计表达的，因此到底损失多少还真不一定。

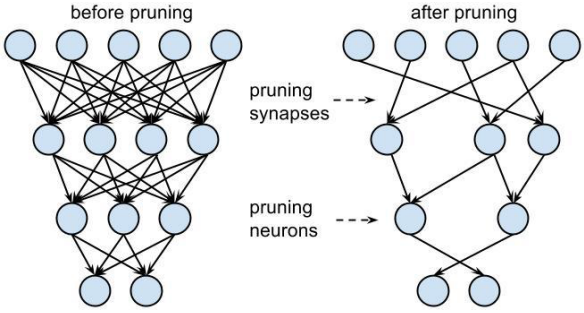


图 12.10

以上是 CNN 的小结。不要以为图像处理与你无关，其实一篇文章如果你把它转化为一个矩阵，无非是一个图像；一段音频你给它转换成一个矩阵，也无非是一个图像，你看，都可以和 CNN 挂钩。

13 时间序列与 RNN

时间序列，是一个随时间变化的过程，如果把它像一个数列一样排列下来，序列里的数字往往看起来在随机波动。一定程度上，我们可以把它看作一个一维的图像或向量，这个图像不停地向前滚动。

首先这样一个结构可不可以还用之前的 CNN 呢？答案是都行。为什么可以用 CNN？因为可以当成一维的图像，所以当然可以使用图像的一些方法。为什么不可用 CNN？因为其依赖的基本假设是不一样的。笔者的总结就是看时间这个概念在你的预测里有多重要。图像，我们通常不把它看作一个不停向前滚动的东西，而有些东西，比如股市大盘指数，是一个明显的动态过程。当然，如果换一个角度，比如股市的 K 线图，那么也可以把时间序列看作静态的图像。而如果关心在一个图像里一个人是如何从左扫到右的，不同的扫描方式又能得到什么结果？那么图像也能看作一个动态过程。因此，是图还是序列不重要，重要的是动和静。

RNN，就是一个处理动态过程的网络，CNN 擅长处理静态的关联。用一句话总结，RNN 具备的素质就是把历史的输入合理地转化为此刻决策用的特征，也可以称之为记忆力。比如，我要决定今天看不看电影，可能取决于我一小时前做没做完工作，也可能取决于我昨天看没看过电影，我们要把这些因素综合起来，加上此刻的状态（比如天气好不好），一起构成决策因素。

首先，当我们的输入是一个序列时，如何能够保留输入序列的历史供我此刻的决策使用呢？对于序列输入，最简单的方法就是构建一个从过去所有输入到这一刻输出的神经元的全连接网络，每一个过去的点到此刻的影响是 w，我们用 sum（wtxt）表示这个积分。当然我们无法保留过去所有的信息，所以 w 一般随着时间减少，最后在某个点截断，代表记忆的尺度。这个非常简单的模型构成了最简易的线性时间序列分析器。但有三个很核心的问题没有解决：

（1）我们需要指定相等长度的记忆窗口。

（2）需要学习的参数太多，如果一篇非常长的文章是输入序列，则是无法接受的。

（3）不能刻画过去对于现在的非线性影响，一切过去对于现在的影响都是线性加和，而忽略了可能的非线性效应，这使得对“记忆效应”的刻画非常有限。这一点可以通过加入核函数的方法缓解。

那么，是否可以使用和 CNN 一样的技巧，把整个输入切分成小块，用一个卷积核把它们卷到一个隐层网络里，最后通过读出网络，得到整个历史的信息？可以！如果这个隐层比较大，我们还可以省掉那个过去到现在的全连接网络，而是只把现在的信息输入隐层，隐层把它缓存起来，不是马上忘掉，这样就得到了一个更优雅的结构。

这个结构，非但优雅，而且有效。这个结构能够起作用，其核心依然在于我们放入了强大的预设。假设的第一条是，此刻的决策，与最近的历史关系最大，而较远一点的次之，即使与较远的历史有联系，它也是通过较近的历史来传递的，这和人脑的工作记忆原理有些相近。而且过去对现在的影响，在时间轴的每一点上都一样（表现在从 $h(t\text{-}1)$、$x(t\text{-}1)$到 $y(t)$稳定的概率分布），这里与卷积网络的权值共享道理是一样的，姑且叫作时间轴上的平移不变性。上面所说的假设如果成立，RNN 就会是一个十分有效的工具。

然而我们很快遇到一个问题，一位的数据不够存，毕竟，计算机的缓存还要 3MB 呢。至此，我们把隐层拓展成一个互相连接的神经网络，由连接矩阵 $\boldsymbol{W}$ 刻画。你可以理解为信息被输入一个隐层后，连接输入的神经元会向其他神经元喊话，这样神经元之间不停地喊话，会产生一种非常特殊的回声效应，如同信息在网络里回荡，“循

环”两个字，正点出了这种特征，如图 13.1 所示。

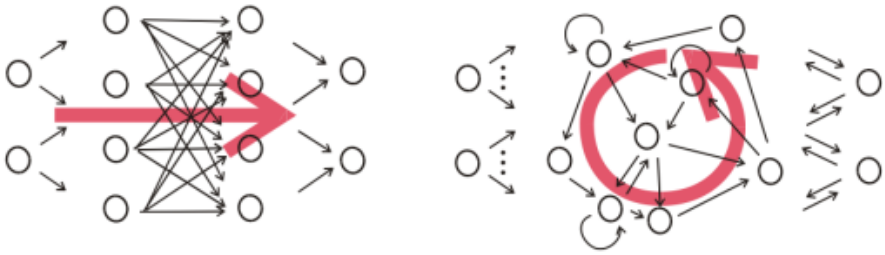


图 13.1

RNN 的动力学本质

如果把对 RNN 的理解再提升一个层次，就需要把之前的非线性动力学抬出来。说隐层内的神经连接的作用，正是刻画动力学的本质——反馈。循环正对应动力学系统的反馈概念，可以刻画复杂的历史依赖：路径依赖。另一个角度看也符合著名的图灵机原理。即此刻的状态包含上一刻的历史，又是下一刻变化的依据。这其实包含了可编程神经网络的核心概念，即你有一个未知的过程，但你可以测量到输入和输出，假设当这个过程通过 RNN 的时候，它是可以自己学会这样的输入输出规律的，而且因此具有预测能力。从这点上说，RNN 是图灵完备的。

我们用一个非常简单的例子说明 RNN 是如何用非线性动力系统给问题编码的。假设我们有三个二进制数，然后有一个输入输出表格，最简单的情况是要执行一个 flip-flop 操作，也就是说，要做的是记录输入的变化，当输入和当下输出不一致时，输入就变到和输出一致，其余操作不变，如图 13.2 所示。神经网络如何做呢？很简单，我的 2^3 中输入，我可以用神经网络系统的 8 个定点（Fix Point）给它编码，每个定点对应一个输入输出的情况。然后，当输出状态改变时，就是从一个定点跃迁到了下一个定点（中间经过鞍点）。

如果是对于大量含有振动的时间序列呢？没关系，我们还有各类 Limit Cycle（极限环）和 Hopf Bifurcation（霍普夫分岔）在那里守候。

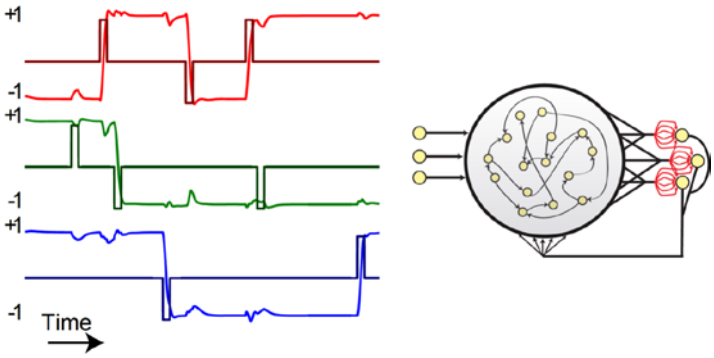


图 13.2

RNN 甚至可以用来模拟看起来是混沌的一些系统的动力特性，了解其时间序列的一些属性。

总的来说，作为一个非线性动力系统的 RNN，可以继承系统内的非线性的非线性。即使是一个内在不可预测的时间序列，它也可以通过对定点、环和混沌的把握，把时间序列的一些本质学起来，最终起到一个万能动力系统逼近的作用。一旦 RNN 学习到了真实数据背后的动力系统的性质，它也就掌握了过程中复杂的路径依赖，从而能够对过去和现在进行建模。

让我们来看看 RNN 的具体玩法，如图 13.3 所示。

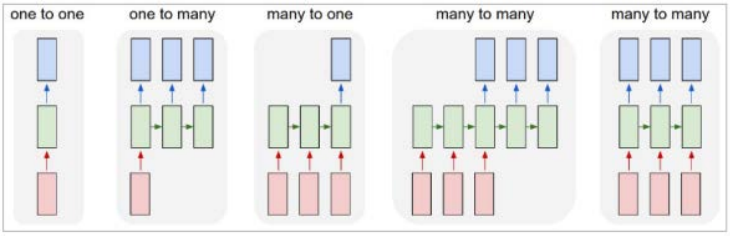


图 13.3

图 13.3 中第 1 个图是 CNN 的架构，第 2 个图到第 5 个图是 RNN 的几种基本玩法。第 2 个图是把单一输入转化为序列输出，例如把图像转化成一行文字。第 3 个图是把序列输入转化为单个输出，比如情感测试，测量一段话正面或负面的情绪。第 4 个图是把序列转化为序列，最典型的是机器翻译，注意输入和输出的“时差”。第 5 个图是无时差的序列到序列转化，比如给一个录像中的每一帧贴标签。

图 13.4 中的 ***h*** 就是 Hidden Variable（隐变量），即整个网络每个神经元的状态，***x*** 是输入，***y*** 是输出，注意，这三者都是高维向量。隐变量 ***h***，就是通常说的神经网络本体，也正是循环得以实现的基础，因为它如同一个可以储存无穷历史信息（理论上）的水库，一方面通过输入矩阵 ***W_xh*** 吸收输入序列 *x* 的当下值，另一方面通过网络连接 ***W_hh*** 进行内部神经元间的相互作用（网络效应，信息传递），因为其网络的状态和输入的整个过去历史有关，最终的输出是两部分加在一起共同通过非线性函数 tanh。整个过程就是一个循环神经网络“循环”的过程。***W_hh*** 理论上可以刻画输入的整个历史对于最终输出的任何反馈形式，从而刻画序列内部，或序列之间的时间关联，这是 RNN 之所以强大的关键。

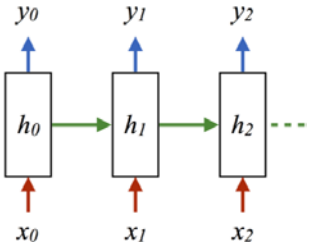


图 13.4

RNN 的本质是一个数据推断（inference）机器，只要数据足够多，就可以得到从 *x*（*t*）到 *y*（*t*）的概率分布函数，寻找到两个时间序列之间的关联，从而达到推断和预测的目的。这里我们无疑回想到另一个做时间序列推断的神器：HMM（隐马尔科夫模型）。在这个模型里，也有一个输入 ***x***、输出 ***y*** 及一个隐变量 ***h***，这个 ***h*** 和刚刚的 RNN 里的 ***h*** 的区别是迭代法则，隐马尔科夫模型通过跃迁矩阵把此刻的 ***h*** 和下一刻的 ***h*** 联系在一起。跃迁矩阵随时间变化，而 RNN 中没有跃迁矩阵的概念，取而代之的是神经元之间的连接矩阵。HMM 本质是一个贝叶斯网络，因此每个节点都是有实际含义的，而 RNN 中的神经元只是信息流动的枢纽而已，并无实际对应含义。两者有千丝万缕的联系，首先，隐马尔科夫模型能干的活 RNN 几乎都可以做，比如语言模型，但是 RNN 的维度会更高。在这些任务中，RNN 事实上是用它的网络表达了隐马尔科夫模型的跃迁矩阵。在训练方法上，隐马尔科夫模型可以通过类似 EM 来自最大后验概率的算法得出隐变量和跃迁矩阵最可能的值。而 RNN 可以通过一般的梯度回传算法训练。

14 会遗忘的神经网络

（阅读难度★★★）

RNN 虽然能够模拟任何一个动力过程，但实际上并没不那么容易使用。RNN 的强大功能，体现在能够学习过去时间点对现在的影响，而这依赖的是 BPTT 回传算法。在回传的过程中，*t-n* 时刻到 *t* 时刻的联系是依靠 *h* 连接的，如果我们暂时不看非线性的部分，那么它所对应的是一个矩阵连续相乘的过程。

想象一下，如果做一个特征变换，我们会得到一系列特征值相乘的情况；如果特征值小于 1，这一系列的特征值相乘得到的结果是 0。对应这种状态的动力学状态事实上是稳定的，即之前讲过的定点。反过来呢？也就是特征值大于 1 的情形，我们会得到过去的信息在链式法则里被连续放大，对应之下，即过去任何一个微小影响都会使得整个系统爆炸。如同之前讲过的混沌系统，任意一个初始值的微小波动都会导致结果的一场风暴，这样的训练结果显然是我们无法接受的。如此看来，RNN 只能生活在混沌和稳定的边缘，即特征值比较接近 1 的状态。

前面提到的就是 RNN 训练里著名的梯度消失和爆炸问题，RNN 后来的各种变版，都是用来应对这一问题的。

我们的第一个策略其实是偷懒，既然 RNN 这么难，那么干脆不训练隐层到隐层

的连接 Whh，而是直接训练隐层到输出的连接 Why，这样我们就能得到一个新的模型，叫作 Echo State Network。这个模型的 intuition（自觉）也很有意思，它说的是这个没有训练的隐层神经网络，好比一个大森林，过去的输入好比一个人留在森林里的脚印，这些脚印虽然踪迹模糊，但是此刻，我们只要进入森林，就有足量的信息把握整个人的历史，预测人的行踪，你只需足够聪明地读取这些信息即可。从另一个角度看，有一点像 SVM 里的 Kernel Machine，这个东西把输入的信息通过一系列非线性变换导入一个高维的非线性空间，反正需要的东西都有，你只要聪明地读取就好了。

当然，这里的关键是我们要有足量的重构过去的信息，此中要害就是刚刚讲的让大部分的神经元处于这个稳定和混沌的交接状态。

Echo State Machine 已经可以做到不少时间序列的预测，但无法训练隐层，所以，我们抛出了最终更加广泛使用的一个解决梯度消失和爆炸的工具——LSTM（Long Short Term Memory）。很多人说 LSTM 是增加了记忆的 RNN，实际上是一种误读，RNN 理论上可以存储任何长度的信息，但实际上不能，而 LSTM 通过增加控制机构，使得这个变成可能。

之前导致梯度消失的原因在于 Whh 的特征值连续相乘导致的衰变，那么可不可以这样做，也就是说，让每个时刻的特征值都不同，在我们需要它衰减的时候它衰减，不需要它衰减的时候就不衰减？

看似可以，但是模型会变得很复杂，失去了之前说的权重共享的美妙特性，但是，我们可以进一步引入假设，即尽量保持 RNN 的简洁，虽然特征值可以随时间变化，但是如何变化依然是由每个时刻的隐层和输入的状态决定的，因此我们引入门电路的概念。

门就是控制信息衰减（或者新的信息和旧的信息的比例，遗忘和更新的比例）的阀门，这个阀门由输入和隐层状态决定，只要控制好这个阀门，我们就可以让梯度衰减（反过来信息消失的时间尺度）控制在一个合适的范围内，而且这个范围是随着输入动态调整的。信息闸门的原理是动态地在遗忘和记忆之间调整自身。

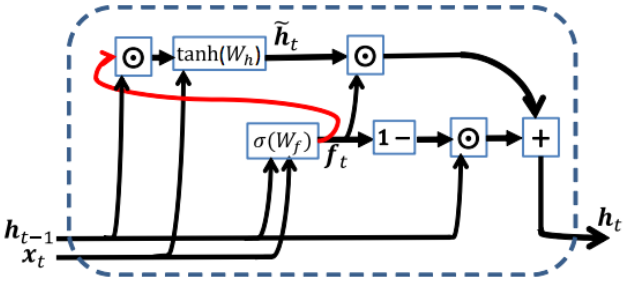
这样我们就得到了 LSTM 的完整表达。

这里首先增加一个隐变量作为记忆单元，然后把之前一层的神经网络再增加三

层，分别是输入门、输出门和遗忘门，这三层门就如同信息的闸门，控制先前网络内被保留的信息数量和新进入的信息数量，而且门的形式都是可微分的 Sigmoid 函数，确保可以通过训练得到最佳参数。

根据这一原理，我们可以抓住本质，简化 LSTM，如 GRU 或极小 GRU。其实我们只需要理解这个模型就够了，它们甚至比 LSTM 更快、更好。

下面看一下最小 GRU 的结构，如图 14.1 所示。



$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right), \quad (6a)$$

$$\tilde{h}_t = \tanh\left(W_h\left[f_t \odot h_{t-1}, x_t\right] + b_h\right), \quad (6b)$$

$$h_t = (1 - f_t) \odot h_{t-1} + f_t \odot \tilde{h}_t. \quad (6c)$$

图 14.1

第一个方程 f 是遗忘门；第二个方程与先前的 RNN 对比，它们的结构相同，只是让遗忘门 f 来控制每个神经元放多少之前的信息出去（改变其他神经元的状态）；第三个方程描述“惯性”，即最终每个神经元保持多少之前的值，更新多少。

理解了这个结构，也就理解了记忆体 RNN 的精髓。

下面看一下 Google 翻译是怎么做的。翻译是沟通两种不同的语言，而沟通的本质是因为它们所表达的事物是相同的，我们自己的大脑做翻译的时候，也是根据它们所表达的概念相同来沟通两种语言的（比如苹果 vs apple）。如果汉语是输入，英语是输出，则神经网络事实上做的是下面这样一件事。

Encoding：用一个 LSTM 把汉语变成神经代码。

Decoding：用另一个 LSTM 把神经代码转化为英文。

第一个 LSTM 的输出是第二个 LSTM 的输入，两个网络用大量语料训练好即可。Google 这一次是在其中加入了 Attention 机制，这样 Google 的翻译系统就更接近人脑了，如图 14.2 所示。

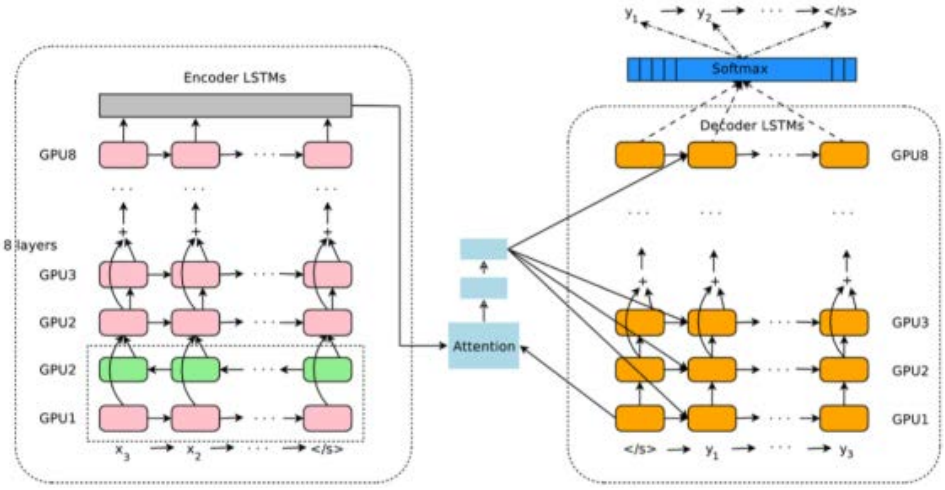


图 14.2

运用记忆神经网络翻译的核心优势是我们可以灵活地结合语境，实现句子到句子，段落到段落的过渡，因为记忆特性使得网络可以结合不同时间尺度的信息，而并非只抓住个别单词，这就好像你能够抓住语境而非只是望文生义。也正因为如此，RNN 有着无穷无尽的应用想象力。

15 跟着 AlphaGo 理解深度强化学习框架

（阅读难度★★★）

AlphaGo 这个被大家热议过很久的主题，里面其实包含了非常深刻的机器学习智慧，是学习机器学习甚至是人类决策的绝好教材。机器学习是 AlphaGo 取胜的关键，为什么它会发挥巨大作用，请看下文。

了解一门科学技术最好的方法就是找出其核心论文，让我们看看 AlphaGo 的核心论文是怎么解读这个问题的。如果把你放在这样一个位置，你会如何设计这盘游戏。

如果大家了解棋牌类游戏以及电脑与之对弈的历史，就会非常清楚老派程序员的套路，也就会明白这类问题最简单的办法就是穷举法，比如历史上著名的八皇后问题，你需要在国际象棋棋盘上摆放八个皇后，使得她们各自不位于对方的纵线、横线或对角线上，你只需要按照一定的方法做一个循环，从第一排往下一排遍历，当碰见摆不开的情形时，就回到上一步重新摆，最后总可以摆出来，如图 15.1 所示。

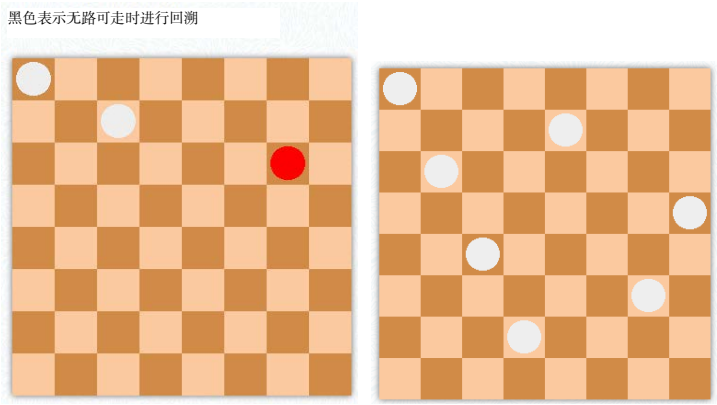


图 15.1

与之类似的方法稍作改进可以很好地解决国际象棋的问题，却难以解决围棋的问题，为什么？因为众所周知，围棋的维度实在太大了，每一次落子都有几百（棋盘 19×19 大小）种可能。设想假如一盘棋要在几百步之后得出胜负，会有多少种可能性，确实很难通过任何和穷举法沾边的算法解决。

这里就涉及如何有效地减少搜索空间这个核心问题。这也是为什么一个下围棋的问题需要用到机器学习，因为机器学习可以让你通过有限数据推测所有其他可能（类似一个插值过程）。

在让机器做这件事之前先看看人是怎么做的。我们无时无刻不在决策，也面临如何减少搜索空间的问题。虽然人生有无限种可能，但大多数可能你连考虑都不会考虑。我们人类用愚蠢和聪明、合理与不合理这些词汇描述各种选择的优劣，并且大脑自动屏蔽大部分不合理的解释。你是如何得到这些答案的呢？第一个就是通过常年的试错来计算每个行为的结果，所谓一朝被蛇咬，十年怕井绳。另一个就是看书，和高手对话，直接学习他们的经验。

反过来就是机器学习的原理，首先说试错学习，或者根据某种行为最终导致的结果来调整行为策略的方法，我们通常称之为强化学习。

强化学习如图 15.2 所示，Agent 会根据环境给予的 reward 调整 action 的一个反馈系统，最终实现利益最大化，难点在于 Agent 的行为通常会改变环境，而环境又会影响行为策略。

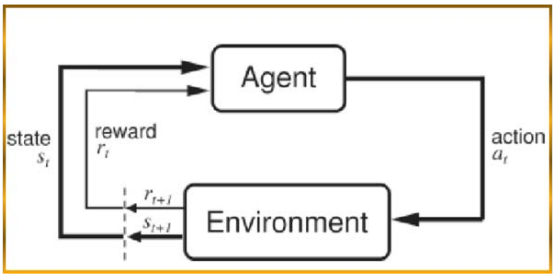


图 15.2

具体到围棋上，这个策略的核心是根据围棋的特性：

（1）在每一步双方信息完全已知；

（2）每一步的策略只需考虑这一步的状态。

这允许机器学习用一个非常凶猛的简化框架来解决这个问题，即马尔科夫决策过程。也就是说，我们用一个离散的时间序列来表述状态 s，用另一个离散的时间序列表述行为 a，两个时间序列有着深刻的耦合关系，下一刻的状态 $s(t+1)$取决于此刻行为 $a(t)$和状态 $s(t)$，最终决定下一刻的行为 $a(t+1)$。两者间的关系即策略 $P(a(t)|s(t))$，由于是马尔科夫链，所以每一时刻的策略只与此刻状态 $s(t)$有关。各种棋类就是最明显的马尔科夫链。由于未来存在不确定性，因此策略本身也是一个概率分布函数的形式。最终我们要优化，使得 $P(s|a)$所得到的回报 $R(s)$最大。马尔科夫决策过程是在解决未来状态不确定而行为和状态又具有马氏性时十分有利的方法。

解决马尔科夫决策过程的一个简单实用的算法叫作蒙特卡洛树搜索（MCTS），如图 15.3 所示。

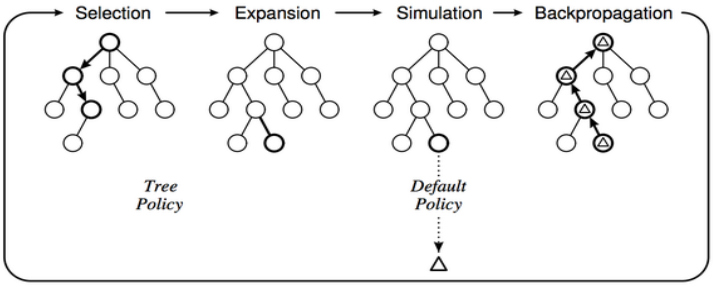


图 15.3

图 15.3 描述了蒙特卡洛树与它的四个步骤：选择、扩张、模拟估值和结果回传，对应一个经典的强化学习框架。

蒙特卡洛是大名鼎鼎的随机抽样方法。提到树，大家一定可以想到决策树，树的节点是某一刻的状态，枝杈代表一个决策。而这里的蒙特卡洛树，就是用随机抽样的方法生成整个决策树的过程。

假设电脑现在的状态是 $s(t)$，那么你随便扔个骰子走一步，然后电脑模拟的对手也扔个骰子随便走一步，这样下去，总有一刻会分出胜负，这个时候你回顾胜利和失败的人的历史走棋轨迹，赢的走法在其整个决策树上的每个状态（枝叶）都加一分，输的走法每一步位置都减一分，这个分数会影响下一次抽样的概率，使得容易赢的步子会有更大概率取到。玩无数次后，就会选择出特别容易赢的策略。这个过程酷似进化选择算法，就是让那些有优势的选择有更高的繁殖子代概率，从而最终胜出，体现了生物和环境的博弈。

> 注：这里提个小问题，马尔科夫决策过程威力巨大，而在现实决策里，你要去哪个大学上学，你要选哪个股票，都很难只看当下，这个框架还能用多少呢？

以蒙特卡洛树为代表的强化学习在围棋这种走法可能性超多的情况下，只能部分地减少搜索空间，使得电脑达到一个高级业余选手的水平，而如果要进一步减少搜索空间，应该怎么办呢？人类减少搜索空间的一个重要方法是学习高手经验，背棋谱，看得多了，就有一种犀利的直觉走出一个妙招。转化为数学语言，就是通过看棋谱，取得一个在某种局面下任意策略和最终赢率的对应关系，即使这个局面你从未见过。

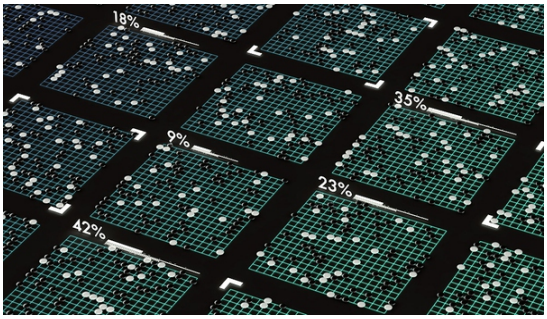
> 注：我们经常会觉得直觉这个东西好像是天上掉下来的异禀，实则恰恰相反，直觉才是最需要学习的。

让机器来做就是有监督学习的回归算法，你要提取棋局的特征，算出对应每一个走法出现的概率 $P(a(t)|s(t))$，然而围棋棋局的特征实在太复杂，这时候我们的深度学习开始派上用场，它可以自发地学习事物的表征。

机器学习训练的目标是使得数据被观测到的概率最大，所谓 Maximum Likelihood，对于神经网络，就是网络连接参数的调整。

深度学习的过程正如同我们见识一个东西多了，自发地开始具有举一反三的能力，自然而然地把直觉加入了策略选择，这时候你可以通过有限的经验把握无限。在训练过程中，AlphaGo 不停地根据现有的局面预测专家可能会出的招，在经过三千万组数据的训练后，深度学习可以达到 55.7%的预测率，这个概率说明人类的意图也并不难被猜中，这也是很多人说和 AlphaGo 下棋如同和无数高手过招的原因。当然，这还不是训练的终结，此处的神经网络只是在描摹高手的动作，而之后我们要让它能赢，好比在实践中理解和优化高手的招术，这就是训练的第二步，用强化学习方法，训练网络连接系数，具体方法是让现有的策略网络和随机选出的一个之前的策略网络进行左右互搏，然后把胜负结果回传到每一步的策略上，进行梯度训练。经过这个过程，策略网络可以成功战胜一些中级爱好者水平的算法和自己之前在描摹各种高手时候的状态。

策略网络的思维是计算每种走法出现的概率，如图 15.4 所示。



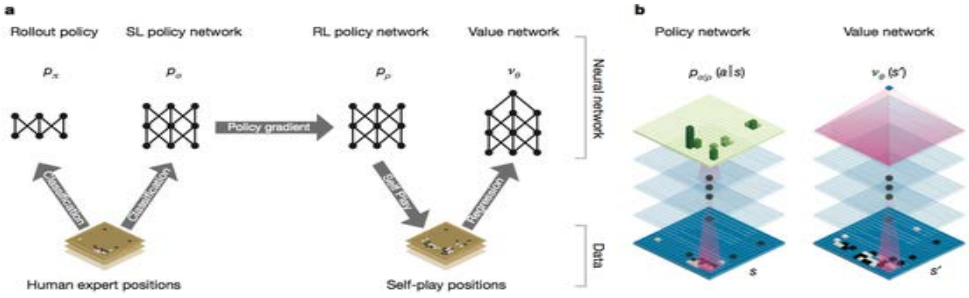


图 15.4

训练的最后一步是估值网络。估值网络是做什么的呢？首先，在一个强化学习框架下，你需要知道每个行为所对应的确定回报，难点在于围棋是下完棋才有确定回报的，想想围棋步骤中的无限多可能性以及得到结果可能的步数就令人生畏，此处深度学习算法的作用正是不需要走完就巧妙地估计出这一步对应的盈利期望，过程需要用一个深度网络通过强化学习的框架来进行。估值网络的本质在于建立现有行为和长远收益的联系，有人称之为看趋势和全局观。

公式如下，训练要解决的问题是，求得状态 S 下采取策略 p 的最终收益。

$$v^p(s) = \mathbb{E}[z_t | s_t = s,\ a_{t\ldots T} \sim p]$$

估值网络的效果图如图 15.5 所示。

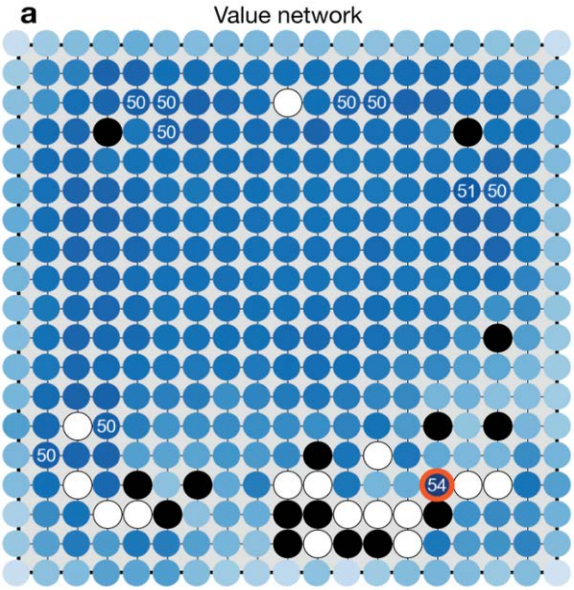


图 15.5

那么问题来了，蒙特卡洛树和深度学习两者是如何天衣无缝地结合起来的呢？这就是整个 AlphaGo 设计最巧妙的地方。首先蒙特卡洛树可以拆解为 4 步：

第一步，Selection，在已有的选项（经历过的）中进行抽样选择。

第二步，Expansion，走到一个先前从未经历的局面上，探索新行为，即生成新的枝杈。

第三步，Evaluation，得到新行为的回报。

第四步，Backup，把回报的结果反向传递给策略。深度学习的结果可以被非常完美地嵌入蒙特卡洛搜索的步骤里，首先，在 Expansion 的阶段，我们不用从零开始随机生成一个前所未有的状态，而是根据前人经验训练的策略网络直接生成新状态，海量减小了无用的搜索。然后，在 Evaluation 步骤，我们无须跑完整个比赛，而是通过深度学习的结果直接算出这个新举措可能的回报（此处即估值网络的作用），这个计算出的回报，会在最终游戏完成的时候与真正的结果相结合，从而完成学习的步骤。

深度学习嵌入蒙特卡洛树搜索的方法如图 15.6 所示。

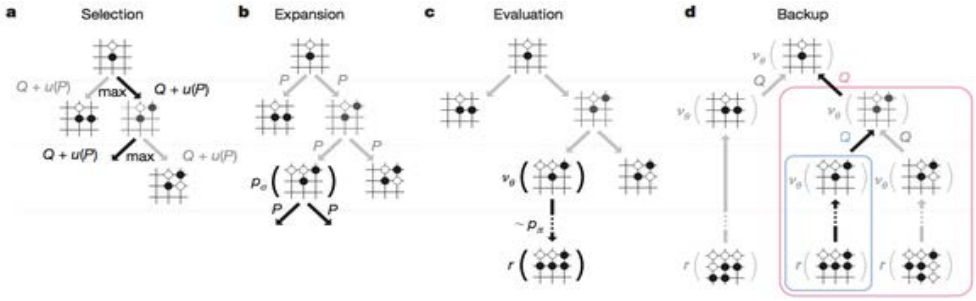


图 15.6

与战胜国际象棋大师的深蓝不同，在 AlphaGo 这里，机器学习发挥了巨大的作用，因为 AlphaGo 的策略和智能主要是在不停地看棋谱和左右互搏中进化出来的，对于围棋这样规则非常复杂的东西，设计一套必胜规则几无可能，只有机器学习（强化学习）的进化和自我改进思想才是最终取胜的法器。因此 AlphaGo 的技术对其他人工智能非常有启发。

从整个解析来看，其实训练 AlphaGo 的算法思路并不十分复杂，用一句话总结，就是站在巨人的肩膀上迅速试错。这可能也是各种人生决策的最好方法吧。然而，我们人类没有那么多时间玩 Simulation，也没有那么多 GPU 进行并行运算，所以我们其实寻找的是低搜索成本的近似解，谓之次优解。

16 从阿尔法元看强化学习的更广阔潜力

2017 年 10 月 19 日，《自然》杂志报道，谷歌新智能阿尔法元（Alphago Zero）超越自己的大哥——AlphaGo，在无任何人类指导的情况下，自学围棋，并以 100：0 的战绩战胜 AlphaGo，“元”含有起点，创世之意。总之，就是从零开始，其实这个元字用意很深，一方面，这个算法是不需要人类数据指导的，也不需要它哥哥（AlphaGo）指导，就自己演化出来。另一方面也可以理解为它可以开启新纪元。

AlphaGo 已经刷爆了人们的眼球，那么它的弟弟阿尔法元到底有什么能耐，能够再掀狂澜？

这个要点就是强化学习，当年 AlphaGo 出来时，人们说，秒杀人类智商的 AlphaGo 无非是比人类看棋谱的速度快，难道还真的懂围棋吗？它所做的顶多是模仿，里面的强化学习到底有多少作用，真的不知道。那立下首要功勋的深度卷积网络，却需要学习三千万组人类数据进行训练，而整个训练过程需要的能量据说要耗费几吨煤炭，这样的过程比起人的学习效率，还是相形见绌的。

然而今天，阿尔法元却能够在不用那 3000 万数据的前提下完胜 AlphaGo。从人

工智能的技术角度看，这是强化学习的胜利，在不进行监督学习的情况下，就可以达到一个高于人类的境地。

为什么强化学习如此重要？让我们再来看看监督学习和强化学习的区别。

监督学习、强化学习和无监督学习是机器学习的三大框架。从某意义上说，监督学习是给定输入和输出，机器来学习输入和输出的关系。一个好的监督学习算法犹如一个预言家，它能够根据自己之前见过的输入输出关系来预测未知的输入。

强化学习的三要素是状态、行为和环境奖励。强化学习条件下，学习者每一步看到的是它决策的行为结果，然后导致下一步行动，为了最终取得游戏的胜利。用一句话说：强化学习强在决策，监督学习是预言家，强化学习是决策家。

强化学习更像是一个日常决策中的人。我们看到一个老虎，监督学习帮你识别出来它是老虎，可能刚说出来就被老虎吃了。而强化学习告诉你赶紧跑，你可能就活了下来。监督学习让你成为复读机，而强化学习让你成为生物。

强化学习赋予机器以灵魂，监督学习的那些任务反而是在这个宗旨之下产生的。

回到围棋，我们看看强化学习是如何决策的，再来好好理解一些一下“强化”二字。强化的意味是：强化优势经历，反过来，就是弱化劣势经历。当你走了一部棋导致不好结果，之后被选入这一步棋的概率就降低，而导致胜利的选择被不停地强化，直到你每次都沿着最佳路径前进。这听起来很像进化，但与进化的区别是，进化是严酷的客观环境对随机变化的生物的选择，而强化学习里的单元可以通过梯度下降主动调整策略。

既然强化学习那么牛，为什么 AlphaGo 还用监督学习这个拐棍呢？简单来说，强化学习太难了！

强化学习有两大难题

（1）奖励时间的不确定性。今天的努力，可能一天后回报，也可能十年后回报。今天有奖励的事情，明天可能就导致悲剧。对于游戏里的每一次决策，你都无法获得立即反馈，相比监督学习时时可以得到对和错的答案，这个信息实在太弱了，用来指导学习，效果是极慢的（如何利用这个或有或无的信息，强化学习有一系列方法围绕而来，比如 Q-learn）。

（2）探索与收益的平衡难以掌握。有的人一辈子抱残守缺，7 岁玩泥巴，未来就永远玩泥巴。有的人一辈子都在探索不同的方向，但是换来换去最终庸庸碌碌。而只有恰当把握机会，比如说 27 岁前读书去不同国家，27 岁开始认准一个方向成为大佬，30 岁前各种风流倜傥，30 岁选一个知书达理的另一半从一而终，就有可能达到探索收益平衡。强化学习始终面临是探索更多空间，还是开始用现在经验受益的矛盾。

这两点放到围棋这个搜索空间就犹如宇宙星辰的游戏，估计学习时间也要用生物进化的尺度算，然而阿尔法元所用的强化学习算法，号称解决了这个问题。

仔细看，它和它哥哥 AlphaGo 的差别没那么大，只不过这一次的神经网络完全由强化学习训练，和蒙特卡洛树的融合堪称完美。

之前的 AlphaGo 有策略向量和估值网络（都是深度卷积网络）两部分，策略向量负责把棋盘现在的状态转化为可能的行为概率，这个东西被称为策略向量（Policy，是由每个可能的行为概率构成的向量，简称策略向量）；估值则是输入目前的棋盘状态得到最终结果的概率。这两个网络在这里被合成一个巨大的深度残差网络（卷积网络的一种）。

深度卷积网络计算概率如图 16.1 所示。

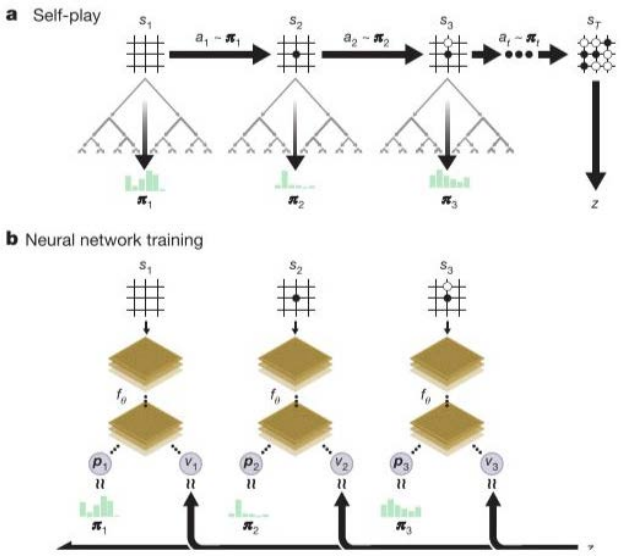


图 16.1

深度卷积网络擅长对图像信息整体进行编码，我们可以把这个巨大的残差网络所做的事情看成白日梦者对未来的总体规划。多层卷积本身的天性决定了它擅长从这种19×19 的格子图像中总结出含义来，强化学习的信息一旦可以训练网络，就会产生意想不到的效果。而之后蒙特卡洛树（MCTS）则对这种初步的结论进行实践（关于蒙特卡洛树的细节可看前文）。

我们来看每一步决策神经网络和蒙特卡洛树是怎样结合的：决策分为搜索阶段和行为阶段。假定现在我处在状态 s，在搜索阶段神经网络对我所能做的所有行为(a)进行白日梦般的预判，即生成赢率 *v* 和策略向量 ***p***(*s*,*a*)。当然这个预判开始很不准确，蒙特卡洛树在此基础上通过无数次模拟实践展开（注意均是在状态 *s* 上），实践出靠谱的策略向量 ***pi***(*s*,*a*)。当然，有了神经网络的帮助，蒙特卡洛树不是随便展开的，也不是从零开始，每一个树的新分支上，我们都通过神经网络给它一个正确步骤的先验概率（*P*）和初始的赢率（*V*），代表走它通向胜利的概率。在神经网络的助攻下，蒙特卡洛树可以更快地更新策略向量（每个行为选择的概率）。此时搜索阶段结束，我们从这个策略向量里通过抽样得到最终进行的行为，即行为阶段。策略更新的方法如图 16.2 所示。

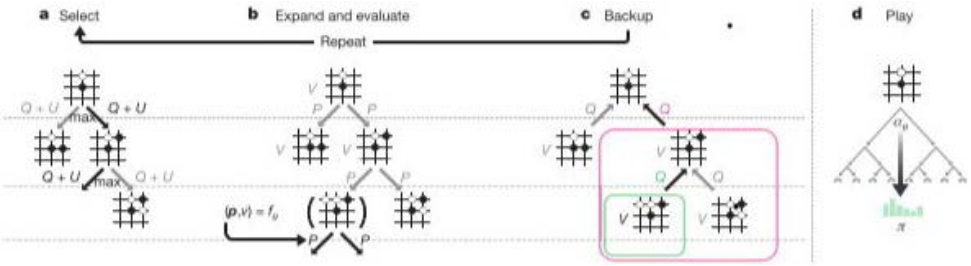


图 16.2

最终当游戏结束时，神经网络的权重开始更新，在更新的过程中，我们把整个游戏的过程分成很多小段，比较神经网络预测的概率和蒙特卡洛树算出来的（策略向量之间的差异）概率，以及预测结果与最终结果的差距，进行梯度下降（梯度由如下公式得到，此处是混合之前的策略向量和估值网络）。

$$(\boldsymbol{p}, v) = f_\theta(s) \text{ and } l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$$

这样周而复始，最终神经网络的预测将越来越精准，和蒙特卡洛树给出的分析越来越一致。而围棋的套路也会被一一发明出来，所谓无师自通。

为什么说阿尔法元敢叫元？如果从技术角度看，这一次的阿尔法元并没有那么多新的东西，而是在之前基础上让强化学习进行的更彻底了，然而它所展示的深度强化学习的应用未来却十分诱人。

首先，并不是每一件机器学习的事情都需要处理数据，尤其是需要大量人力标注的数据，而是可以通过恰当地设立模拟器（比如此处用到的蒙特卡洛树） 来弥补。阿尔法元不是不需要数据，而是数据都是自己模拟产生的。模拟+深度强化学习，在这个简单的游戏规则下，一些复杂的行为范式就可以进化出来，而且可以比人类设计得还好。

这件事在很多设计性的工作里实在是太诱人了。无论是设计新材料、建筑，还是衣服，这些可变维度很高的事物，你都可以想象设立一个模拟仿真环境，再设立一个相应的神经网络去做各种尝试，最终设计出有一个奖惩函数反馈的结果，让这个网络来学习。这就打破了深度学习创业只能和手里有大量数据的垄断者相关的梦魇。

这里的深度强化技术，只展示了冰山一角，在一类被称为 SLAM 的技术上，深度强化学习被证明了强大的控制能力，它能够驱动机器人在非常复杂的空间里进行探索，无须 GPS。这一类深度学习任务，与 AlphaGo 任务不同，因为围棋属于完全信息的博弈，而真正的空间探索，是通过感知系统探测到的不完全信息，通过记忆在时间尺度上的综合，这一点，只能搬出大名鼎鼎的 LSTM 来对付了。

今天，人们已经可以从阿尔法元的妙招里学习围棋，甚至超过人类自身的局限，而不远的明天，说不定我们真的可以在各个人类曾经熟悉的事物上，从深度强化学习形成的策略增进自己本身的智慧。

能够控制运动的深度强化学习，迟早会改变工业界，它不仅是无人车里的核心技术，更是对话、推荐系统、金融交易，甚至是图像识别的利器，几乎各类需要监督学习的事情，说到底强化学习都有实力。你如果制造一个聊天机器人，当然希望它能够揣测你的意图，和你谈天说地而不是背书。你想要的推荐系统，是推荐给你一些你可能想看的书，而不是你刚刚刚看过的书，所以，强化学习才是人工智能的大势所趋。

更有甚者，我们可以设立一个具有类似地球物理环境的地方，让配备了深度强化学习系统的虚拟生物进行各种活动，看它们能否利用这个环境发现和利用其中的物理定律。

第四部分　宇宙间最复杂的就是我们的大脑

17 深层视觉信息的编码机制

（阅读难度★）

认识我们的大脑，就是认识用复杂性对抗复杂性的大自然最伟大的杰作，我们会从中越来越多地观察到和机器学习的联系。

人脑里的 200 亿神经细胞，如同浩瀚的星空。星星交替闪烁，构成我们心智的密码。近一百多年里，这都是科学界最大的秘密。

人们对神经元放电和人的记忆等心理活动关系的最初认识，来自于 20 世纪中期的脑外科医生，神经科学的创始人之一 Penfield 在做一个外科手术时，将一枚电极插入一个女人的大脑里，那个女人忽然看到了 6 年前她分娩时的情景，而且就像真实发生的。这样一个故事，在那个时代算是天外童话一般，以至于那个医生把这种现象归结于：woman is unpredictable——没有人猜得透女人的想法。

从那时起，人类开始明白神经细胞，这种生命系统最神奇的细胞的放电，和我们的内心世界，每一个微小的想法，紧密相连。特定的信息通过神经细胞的放电表达，这个基本过程被称作神经编码，反过来，破译神经细胞的放电求得它所表达的内容的过程称为解码。这两点从根本上说与信息论息息相关。

下面用一个比喻理解编码和解码。为什么叫编码？首先，放电是什么，是神经细

胞间在传递信息，好像一个细胞在向另一个细胞喊话。细胞放电的信号可以被小小的电极测量，这也是我们唯一能够掌握的可测量量。这些电生理学家记录的电信号，就像战争时使用的密码本，说的是有关于我们的所有事情，只是我们看不懂。而解码的过程，就是读懂这个电信号的密码本。

这个解密的过程，有一个阶段是特别容易的，就是视网膜的编码。因为这个装置类似于数码相机，光线通过眼睛这个镜头，投射到视网膜神经元，每个细胞对应相机的一个像素。但是这个相机的美梦止于第一步。当神经信号进入下一层（上丘，视皮层）和更下一层（海马体）时，我们的信心就被打击得无影无踪。神经细胞的信号变得杂乱无序，再也构不成一张照片，如图 17.1 所示。而我们的感觉、心理活动，其实正是来自这些深层脑区。所谓相由心生，我们所说的眼见，其实是这些深层脑区的信息。但是当我们直接看这些信号时，它们却一点不像我们看到的。

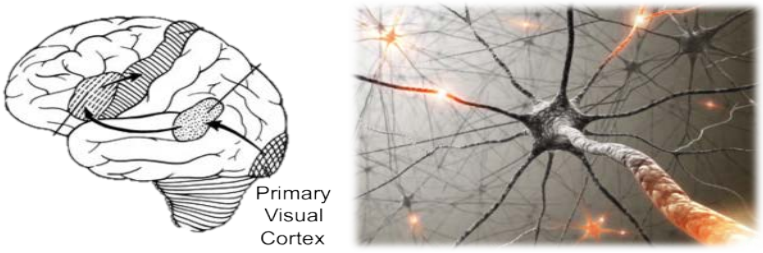


图 17.1

对于视觉编码的研究其实由来已久，在 Quiroga RQ 等人编写的 *Invariant visual representation by single neurons in the human brain.* 这篇文章中指出：深层脑区的编码，指向抽象的概念，而非具体的图画，它就像一个寻找不同事物间不变属性的计算器，通过不变性形成对事物的概念，然后通过概念的联系，重构世界。这个观点与某些古代哲学家（柏拉图和康德）的观点，不谋而合。构成世界基础的是抽象的概念而非具体的事物。

这篇文章首先试图回答认知科学的核心问题：为什么人脑能够在极短时间内辨别一个人或者物体，即使这个人和物体处在不同的状态，甚至穿上了马甲，你都可以给他认出来，而且你可以一下子联系到和这个人吃的饭、说的话等所有细节？

你认出一个人当然是因为你早就认识这个人，负责记忆的核心部位是海马体。海

马体，是一个形似海马的部位。更准确的说法是，它是人脑的 RAM（内存）。因为它负责大脑的情景记忆。这个脑区就像一台小小的摄像机，将发生的事情、时间、地点、人物快速地存入其中，并在较长一段时段保存它（更久远的事情存储在其他部位，这是从一个切除了海马体的病人的情况获知的，它记得童年，却无法形成对昨天的记忆，生活在永恒的“当下”）。

它还有一个另外功能：空间导航（Navigation）。这个脑区的神经元会对视觉信号进行加工，形成对一个地点的印象。这两个功能看似没有关联，实则是一体的。情景记忆无非是时间地点人物，只需在时空坐标上加上对物体的认知就够了。

对事物的任何印象，都要先经过这个脑区。一个人，一张桌子，一幢大楼，都在这个地方编入记忆，而怎么编入记忆，你只需聆听它在说什么。

当然，没有什么人会把它的海马体打开让你插入电极。除了一种情况——癫痫。癫痫是由于神经细胞的大规模同步放电造成的。其实人的神经细胞只有在半死不活的状态下才工作正常，如果它兴奋过度，就癫痫了。当然这是一个痛苦的病患，但是这个病却给读心术提供了绝佳的甚至唯一的机会，因为治疗癫痫的手术需要开颅检测大脑的电信号，确定过度兴奋的区域，这个过程可以持续几天，其间，人可以是清醒的，也就恰好给科学家提供了探索的机会。

某实验纳入了 8 个癫痫病患者，先问他们每个人对什么感兴趣，是某个美女，还是悉尼歌剧院等。然后，当这些人的海马体被插入电极时，科学家就拿出一个大屏幕，给他们看上面出现的某个美女或歌剧院，当然还有路人甲、路人乙，以及一些无关的动物等图片，最后看哪些图像会引起海马神经元的反应（放电）。

如果看到一些细胞对某美女放电了，科学家就再做一次实验。准备第二组图像，是这个美女穿不同衣服的、小时候的、打扮成动物的等图片，甚至还有更抽象的，比如她的名字的符号和录音，看看哪些细胞会反应。

一个惊人的发现是，科学家发现某病人海马体及周边区域内的一组固定的细胞，只要出现和 Jennifer Anniston（美国演员）相关的图像，包括抽象的符号，就会起反应。我们说，这就是这个美女的概念细胞。而这组细胞对不是这个美女的图像表现得兴致全无，几乎都看不到它们放电。示意图如图 17.2 所示。

第一组实验，实验者发现病人脑中一组对 Jennifer Anniston 反应的细胞。无论什么样的 Jennifer Anniston 都会引起这组细胞放电，而其他的物体或人出现时，这组细胞几乎不放电。这组细胞可称为 Jennifer 细胞。

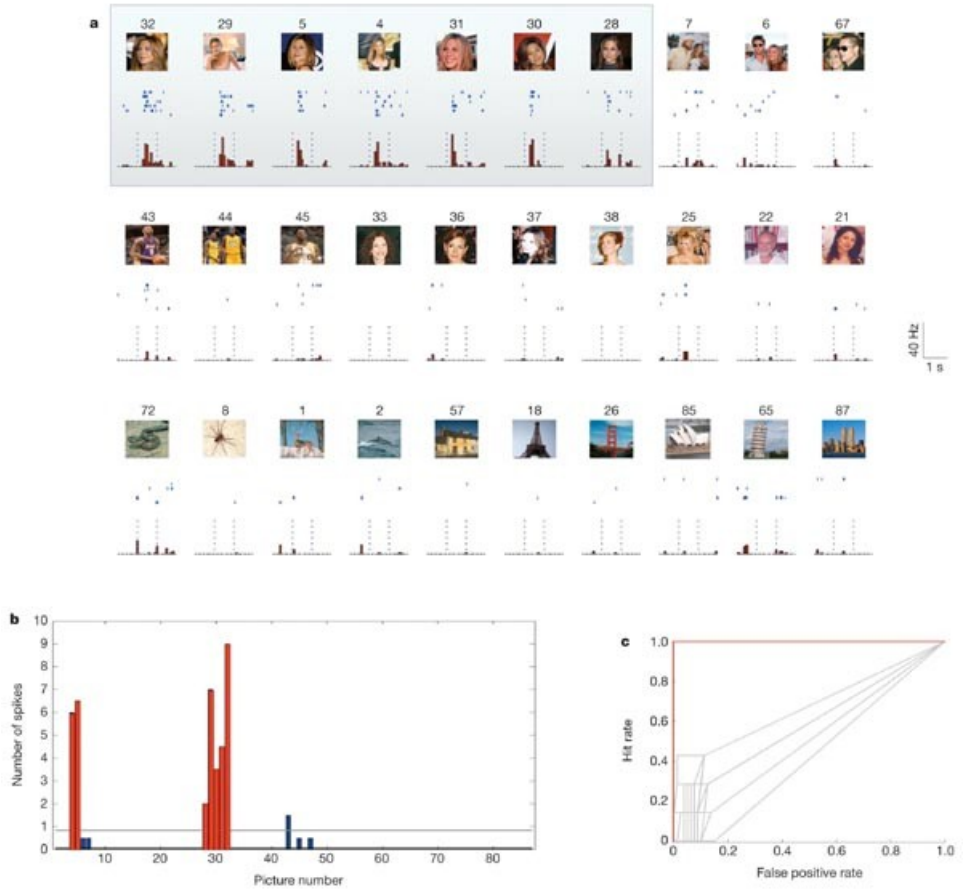


图 17.2

图 17.3 所示的这组是 Halle Berry（第一排），无论是打扮成猫女还是任何发型，都有一组细胞对她敏感，甚至包括她的名字（第一排末尾）这种抽象符号。因此这是实验的第二阶段。Halle Berry 这个概念包括具体的形像、变化的形象，还有抽象的她的符号。

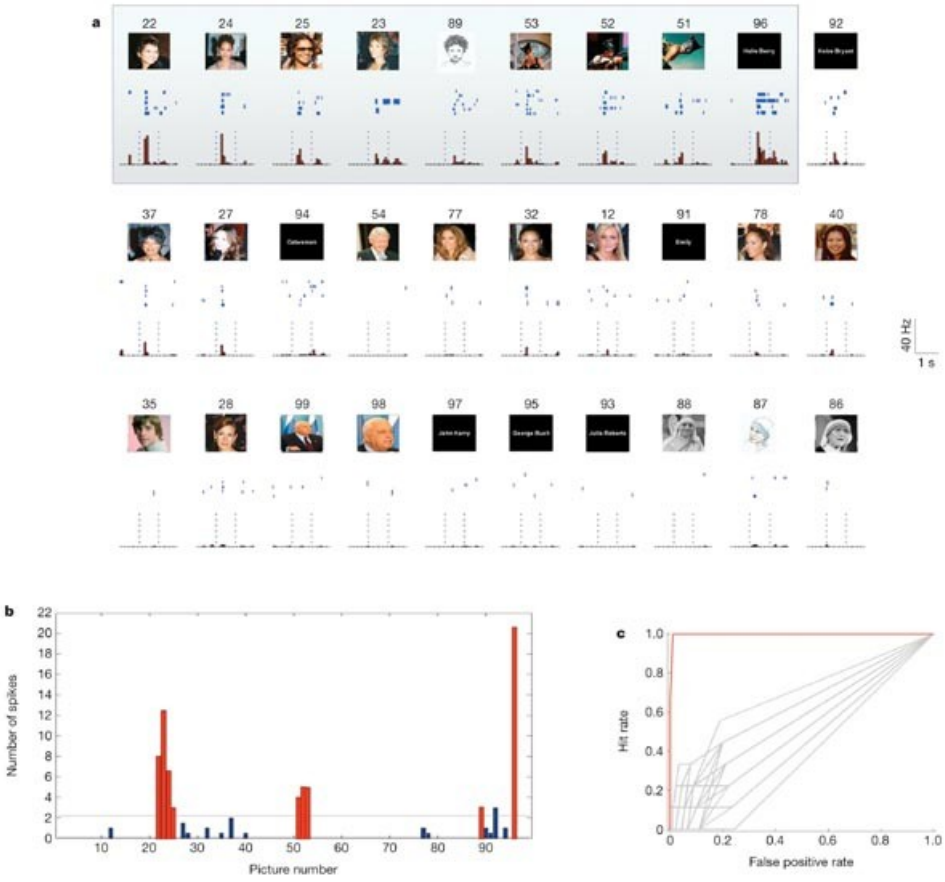


图 17.3

第三组图片把研究拓展到了标志建筑，这是一组对悉尼歌剧院特定放电的细胞，包括歌剧院的形象和它的英文名，如图 17.4 所示。

但是这组实验效果没有上一组明显。

看似貌不惊人的研究，却在提示大脑深层神经细胞的最基本属性：大脑深层神经细胞搜索的是概念，是抽象，越是抽象的，越有效。

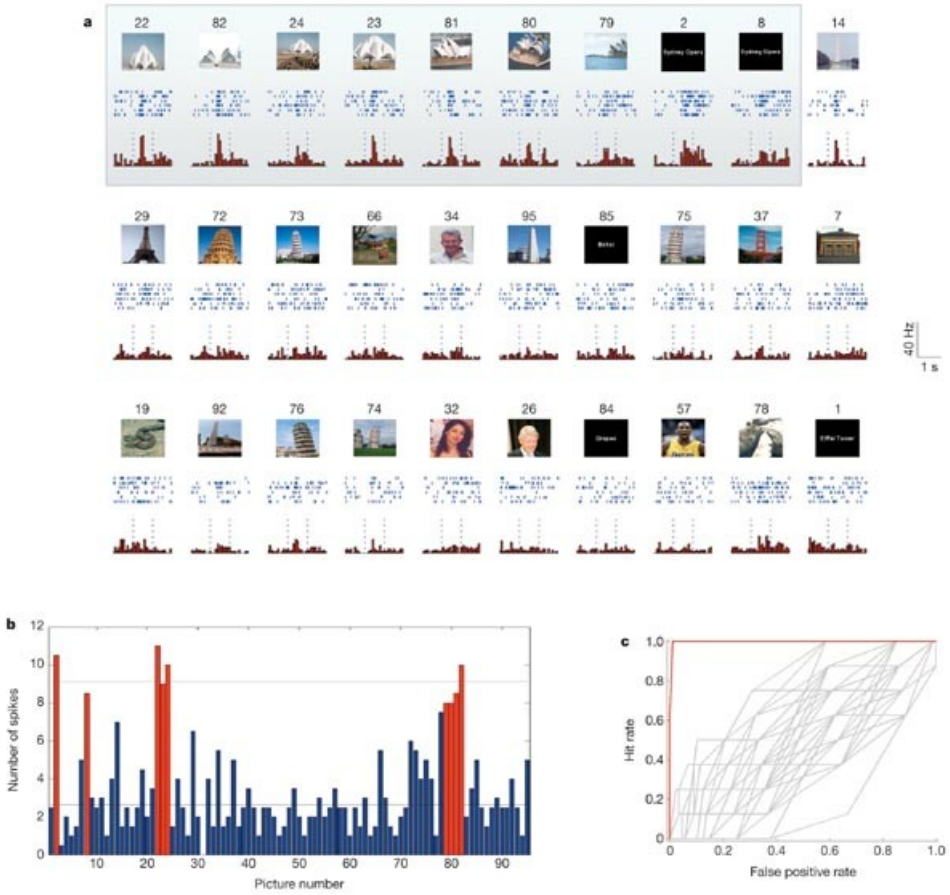


图 17.4

不是说某个大美人的脸吗？不是很具体吗？怎么扯到抽象，扯到概念了？

我们不理解一些东西是因为我们太熟悉。你见到大美女的一张照片，不代表你对她形成了任何认知，她可能只是你眼里稍纵即逝的图像。所谓认知，是一个主动的过程，你从无数的观测角度对事物进行解构，把不同的信息综合。Jenifer Aniston=女人+美貌+名字+年龄+籍贯+她主演的电影，所以你看到，只有对这些抽象的信息解构了，你才对它形成了认知，你把她的信息填入这样那样一个抽象的框框里，好像一个好奇人做的问卷调查。

然后呢？然后你形成了“她”的概念，不是张三，不是李四，而是她，独一无二

的她。独一无二的她是什么？如果她明天长了三头六臂还是她吗？你看，这是一个抽象的问题，就像古代的公孙龙问什么是马，白马是不是马一样。而形成她的概念的关键是，找到不变量。不变，相对什么不变？相对她，即使有一些细节变了，但是你通过对以往关于她大量细节的综合判定，你知道依然是她，因为她不会在短时间内长出三头六臂。具有众多变化中的不变一面，你才得以对她形成认知，形成记忆。

那些大脑深层的神经元，正是通过捕捉纷繁事物间的联系，形成一个个抽象的概念，并通过它们，统领了我们对世界的认知。那些如浩瀚繁星般的神经元，不是具体的一张张脸，而是概念的星空。

其实仔细想想，这也是大自然聪明的表现，如果要像人类设计的电脑一样把一张张图像分作像素保存，以储存你每一秒钟看到的图像，整整一生时间，需要多少台电脑？如果把它们放进一个个抽象的概念，那么一个具体的画面，无非是划过概念空间的一个轨迹，把几个相关概念连接起来就是了。

小明去上学。小明的信息和学校的信息是早已有的，那么在这两个信息之间系个线，就有了小明去上学。

回到起初的问题，海马体是干什么的？空间导航！为什么我们会用这个微妙的空间导航器形成记忆？空间、坐标是什么？不正是一组事物的联系吗（x、y、z）？如果把实物空间变成抽象空间呢？把 x、y、z，变成名字、籍贯、性别呢？是不是很微妙？

这让笔者想到一个节外生枝的问题，为什么抽象的学科（如数学）会如此成功，为什么古希腊的毕达哥拉斯觉得数学的世界才是真实的世界？到底是抽象，还是具体，是人类的幻想，抑或都是？这些问题现在回答还为时尚早。

神经编码，是神经科学最大的谜团，当然不可能只有一种机制，对它们的理解将是人类最大的挑战。但是了解它可能解除的人类痛苦，也将浩大到不可计数。美国、欧盟浩大的脑科学计划的目标，号称人类第二次登月。

18

大脑的自由能假说——兼论认知科学与机器学习

（阅读难度★★）

如果大脑如何运作的机理和热力学一样可以系统完整完美解释，那么这个理论将是物理学家的终极梦想。

而有一篇神文险些胜利，如果你通读此文，也会感觉到一种脑科学、人工智能、物理学，甚至经济学都打通的感觉，然而，估计 99%的物理系读者都很难读通此文（The free-energy principle: a unified brain theory，这套理论可以叫自由能大法）

什么是物理里的自由能法则？总结一句话就是：任何处于平衡状态的自组织系统均趋于自由能极小的状态。这是什么意思呢？自由能又是什么？自由能和大脑有什么关系？

自由能是什么？自由能的物理公式 E-TS 极为简洁，E 是能量，S 是熵。统计物

理说，与外界具备能量交换的系统（一杯放在桌上的热水，底下放着一块冰糖）处于平衡状态下，则自由能最小（水温下降，冰糖扩散），指的是一个能量尽可能小，熵尽可能大的状态。当水温下降到室温，冰糖均匀扩算，此时水分子和外界环境的整体所能够取得的微观状态数最多，也就是最大概然状态，或者说稳态。好比一个教室里没有老师，学生就逐渐变得乱遭遭的，这个乱遭遭的状态就是稳态。自由能最小是热力学第二定律下系统与外界环境相互作用的法则。

生物系统呢？生物系统和水这种系统恰好相反，因为它们是开放的非平衡耗散系统，它们所遵守的法则是趋于有序和结构确定的状态，相当于一个严厉的老师管制下的教室，同学井井有条（这样的状态需要外界输送能量，如老师的工资）。所以我们经常说生物遵守反方向的热力学第二定律，即产生和维持秩序，而紧闭大多数是自由状态。如果某个时刻你开始控制不住自己，那也就是衰老的起点开始。

刚刚好，生物系统也可以被一个形式类型自由能最小的定律结合，表现在认知系统，就是学习过程。

简单来看，你可以把你的脑子想像成刚才说的那杯水，外部环境和这杯水具有一种能量交互的关系，正好对应你的脑子通过眼睛和耳朵采集外部的信息（感知）。这杯水会越来越趋于室温，你的脑子像这杯水一样与外界交换能量和信息，只是，在这个过程中，你的脑子对外界的信息越来越丰富，它不仅是被动地采纳，还要主动地预测和做出行为。

我们的大脑对外界信息的处理过程包含以下几个方面：感知（Ensation）决定计入哪些数据；认知（Cognition）对这些数据进行处理，如分类；行为（Action）是一个由认知的分类导致的决策；一系列的行为组成运动（Motion）。

我们来看看这个过程是怎么发生的。首先，这一切行为的宗旨是为了最大化生存可能性（Maximize Existence）。外部世界充满危险，如果把生物所有的行为可能性做成一个状态空间，那么只有极少数是可以保证生存的，生物越能够绑定在这极少数状态里，生存机会就越大，这也是生物需要抵抗热力学第二定律的理由。那么如何不跑偏呢？

一方面我们有基因编码（Genetic Encoding），使得我么的表现型（Phenotype）

只能在一个有限的范围里（由一个条件概率函数 $P(pIg)$描述），这套编码已经在我们的祖辈自然选择里被遗传下来，去掉了大部分不适宜生存的可能，比如长出三头六臂。

另一方面，作为具有认知能力的动物，能够保证对生存最有利的状态的办法就是学习，整个对大脑外界信号的处理都可以通过“学习”来认识，如图 18.1 所示（请始终想着机器学习）。

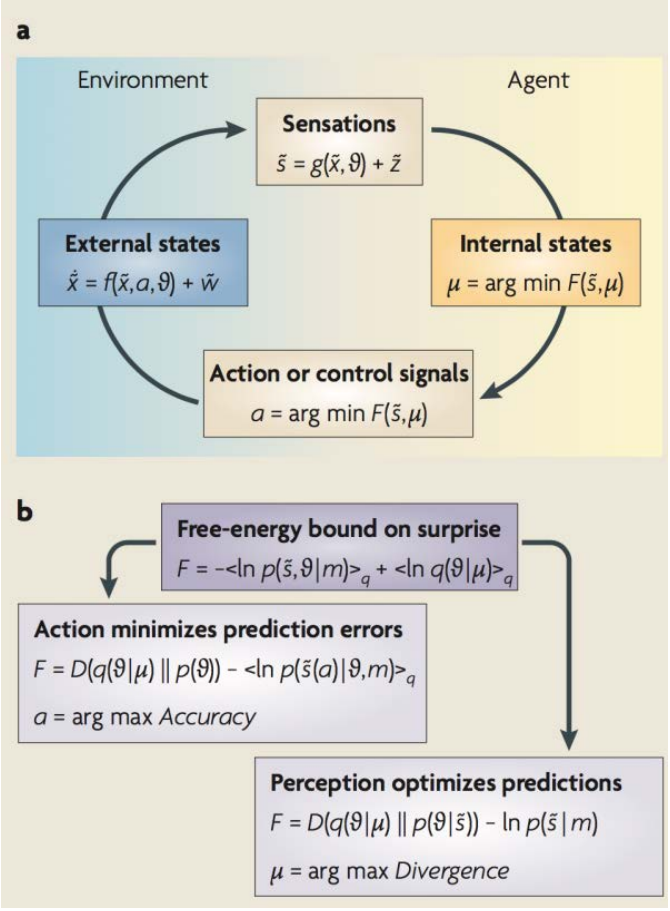


图 18.1

大脑的自由能最小定律说，学习的状态就是使得（1）通过不断调整行为得到符合大脑预期的感知状态（Accuracy），（2）大脑内部的状态能够更加准确地匹配外部世界的变化，不至于出现没有预期到的状况（Divergence）。两部分合在一起使得

图 18.1 定义的大脑的自由能函数最小。这个原则的威力是巨大的，它可以告诉你为什么要这么想、这么看，尽管你平时从未知觉。比如为什么你看到你想看的，为什么你想到的东西又总在随着客观情景调整。

注：第一项中 p 是已知生物的感知得到某种外部世界状态的概率，q 是已知大脑内部状态生物猜测到某外部世界状态的概率，D 是 p 和 q 两个概率分布函数的距离，指的是大脑猜测到真实之间的距离。第二项是已知外部世界的变化得到某种感知状态的概率。

外界的信息是无穷多的，我们的脑子就那么大，选择哪些信息录入，就显得特别重要。感知的过程，就是一个数据采样的过程。做机器学习的读者都知道，开始录入什么样的特征对机器学习具有决定性的一步，而感知也一样，生物要做到的是要取样哪些数据。所以感知绝非被动的过程。

这里的关键是要把感知、认知和行为放在一个循环里理解。这里理解的框架就是贝叶斯决策。贝叶斯公式把事物之间的联系表现为一系列的条件概率关系，并根据新的证据不停调整条件概率，最终优化我们想得到的结果，比如此处的生存机会，就是通常生物系统优化的量，为了优化生存机会，我们只有少数几个想要达到的结果。我们需要产生一个使得这一系列结果机会最大的决策机制（所以我们不是为思考而思考，而是为生存而思考），这一切决定了感知和认知的过程。

那么如何通过学习来找到这个机制呢？其实你会发现，这就是一个模型选择过程（Model Selection），恰恰符合机器学习的本质。

所谓认知，即找到一个预测性模型使得感知得到的信号可以预测出未来外界物体的运动，从而趋利避害。所谓感知，就是寻找最能够提供给这一模型预测效力的有用证据（**Feature Engineering**）。用一句话概括就是，世界很大，我只要最和我的游戏相关的。

那么行为呢？我们的行为最终还是要归之于感官。行为如果带来香甜的巧克力，则得到嘉奖，这是符合你理想的生存预期的。

用一个非常有趣的例子来说明认知模型：一个在黑暗里摸索的人，他不停猜测前

面是什么，它会主动伸手去摸索，来证实他的想法。一般来说，如果我们发现我们的证实与内心的想法不相符，则不停探索，直到符合他的预期，这就是一个简单的认知模型。

这里面我们看到认知模型包含两方面，一个是感知和行为所获取的外部世界的状态，另一个是大脑内部认知过程的内部状态（模型）。这个内部模型不停预测每个一个感官背后的动因和所蕴含的未来变化，而行为本身则趋向于那些有利生存的结果。学习的目的就是让内部状态的模型更加准确（预测精准），另一方面让行为决策获取更多对生存有利的证据。如果模型的预测不准确，则行为决策无法得到正确的结果。

如之前提到的，大脑是部贝叶斯机器，贝叶斯推断和决策的核心即由最新采纳的证据更新先验概率得到后验概率。认知科学的核心（Perception）就是这样一个过程。

这里再说两句认知，认知的过程用机器学习的语言说就是用大脑的内部变量来模拟外部世界，并希望建立内部和外部的一个一一映射关系。因此我们说认知的模型是一个概率模型，并且可以被一系列条件概率所描述。如果用一个形象的比喻来说，你可以把你的大脑看成一个可以自由打隔断的巨大仓库，你要把外部世界不同种类的货物放进不同的隔断，你的大脑内部运作要有一种对外界真实变化的推测演绎能力，即随时根据新的证据调整的能力。你和外界世界的模型匹配得越好，你的脑子的运转就越有效率。认知是对外部世界运动的一种编码，你可以立刻联想到机器学习里的表征方法（Representation），如果你熟悉 RNN 或 CNN 的 Embeding 过程，就会有一种豁然开朗的感觉。

19 大脑中的支持向量机

（阅读难度★★★）

支持向量机（SVM）是一种重要的机器学习分类器，它巧妙地运用非线性变换把低维的特征投影到高维，可以执行比较复杂的分类任务（升维打击）。SWM 看似使用了一个数学上的玄技，实则是恰巧符合了大脑编码的机理，我们可以从 2013 年的一篇论文读起，理解机器学习和大脑工作原理的深层联系（表面的联系是运用机器学习研究大脑）。

论文名称：*The importance of mixed selectivity in complex cognitive tasks*（by Omri Barak al.）

这种惊人的联系可以从哪里看出来呢？首先我们来谈谈神经编码的本质。动物接收到一定信号并根据它做出一定的行为，一个是把外界信号转化为神经电信号，另一个是把神经电信号转化为决策信号，前一个过程叫作编码（Encoding），后一个过程叫作解码（Decoding）。而神经编码的真实目的正是之后解码来做决策。因此，用机器学习的眼光看解码，最简单的方法就是看作一个分类器，甚至是一个 Logistic model 这样的线性分类器，把输入信号根据一定特征分类并分别对待。比如看到老虎要逃跑，看到兔子可以吃掉。当然，有时候解码也在做回归，比如当神经信号最后转化为运动，你需要把神经信号转化为动作幅度的连续变量。好了，这里已经可以明显看到神经编

码和机器学习的联系，神经编码的本质是重新表征信号，从而使得分类或回归容易进行。机器学习的一大类问题本质其实是模仿自然，正如同大多数时候人类如果一件事情做得很好，那往往是仿效了大自然的机制。

下面我们就来看看神经编码是怎样进行的。首先，神经元基本可以看作一个根据外电压调整电阻和电容的 RC 电路，当外信号足够大时，就会导通，否则闭合，通过在一定时间里放电的频率来表征一个信号。而我们谈编码，往往是对时间做一个离散化处理，认为在一个小的时间窗口里，这个放电率是不变的，这样一个神经网络在这个时间窗口里的细胞放电率排在一起就可以看作一个 N 维的向量，N 是神经元的个数。这个 N 维向量，我们姑且叫它编码向量，它可以表达动物看到的图像，或听到的声音，会引起相应的皮层神经网络的响应，即外界信号的表征。注意此处我们先不研究深度网络。

纵轴是细胞，横轴是时间，图 19.1 表现了我们是如何提取神经编码的。

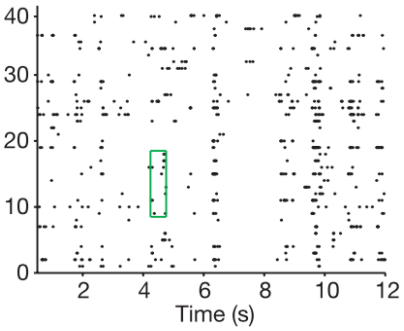


图 19.1

当然 N 维向量和神经编码的真实维度是有区别的，如何定义神经编码的真实维度？首先，我们进入这个 N 维向量所标记的 N 维空间，然后给出所有可能的任务组合。比如给你看一千张图片，假设这些图片代表了整个世界，把每一次我们得到的神经编码标记为这个空间的一个点，最后利用向量代数的思维看这一千个点构成的子空间的维度，即认定为神经表征的真实维度。假设所有的点都在这个 N 维空间的一条线上，那么这个表征是一维的。相应的，如果所有的点都在高维空间的一个二维平面上，则它就是二维的。科学家发现，神经编码的维度通常非常高，当然它不能高于 N，如果神经编码的维度很低，就没有必要用那么多神经元了。

除了编码的真实维度，还有一个概念就是外信号的真实维度，这里的信号是指神经网络所表达的外部信号，当然你要重述外界信号的所有细节那是一个无限的问题，然而我们分类和决策的根据从来都是关键特征，是一个降维的过程，这也是 PCA 的思想。这里我们可以把真实任务里的关键变量看作任务的真实维度。比如你要控制一个手臂的运动，通常只需要控制关节的旋转角度。如果把它看作一个刚体力学问题，则维度大概不会高于 10 个，我们叫它 K。即便是你要分辨人脸这样的问题，问题的维度依然远低于神经元的个数。

那么科学家就面临一个核心问题，为什么要用比真实问题维度高很多的编码维度和神经元个数来解决这个问题？这不是一种浪费吗？

而计算神经科学和机器学习一起告诉我们，神经表征的高维特性正是其所具备的强大学习能力的基础。编码维度越高，学习能力越强。注意，此处我们甚至没有开始涉及深度网络。为什么这么说呢？这里我们说神经编码的机制用到了类似 SVM 的原理，当我们把一个低维度的信号投射到高维时，我们就可以做越多的分类（classification），即使是一个线性的分类器，你也可以解决无数问题。是如何做到的呢？它又是如何与 SVM（支持向量机）原理相通的？

注意，此处讨论的神经编码主要指高级神经中枢的神经编码，比如文中讨论的前额叶（Prefrontal Cortex，PFC），因为低级神经中枢的编码规律并不太涉及分类和决策。

神经编码的奥秘也正是通过神经元个数 N 和真实问题维度 K 的关系（这种差距足可以达到 200 倍）揭示的。为什么看似冗余的神经元个数可以带来质的飞跃？首先，假设当我们的编码维度等于真实任务中关键变量的维度时，我们使用一个线性分类器将不能处理非线性的分类问题（假设你要从西瓜中分离出西瓜子，你不能用一个线性边界把西瓜子从西瓜中剔除出去），这也是在深度学习和 SVM 没有进入机器学习时我们难以解决的典型问题。用 SVM 对这类问题的核心解法被称作重新表征，即把我们的向量从原有坐标系变换到一套新的更高维度的坐标系来表示，这时候我们就可以用分割超平面的方法（依然是线性分类器）来进行模式识别和分类，这样即使西瓜子镶嵌在瓜瓤里，也可以把它摘出去，如图 19.2 所示。

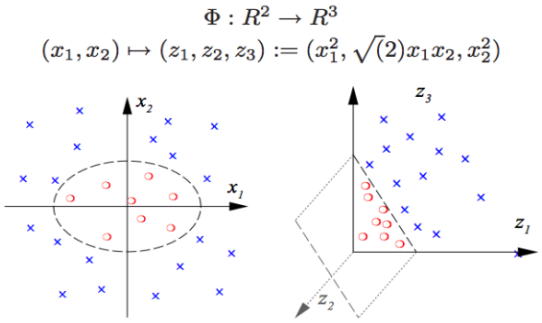


图 19.2

SVM（支持向量机）

SVM 可以进行非线性的分类，例如把图 19.2 中的红点和蓝点隔开，用线性边界我们是无法把红点和蓝点分开的（如图 19.2 左图所示），因此 SVM 用的方法是升高维度。单纯增加变量的个数是不可行的，比如把（*x1*，*x2*）映射到（*x1*，*x2*, *x1+x2*）系统，其实还是二维的线性空间（画成图来看就是红点和蓝点还是在一个平面上），只有使用了非线性函数（$x1^2$, *x1* • *x2*, $x2^2$）我们才有了实质性的低维度到高维度的跨越，这时候就把蓝点抛到了空中，然后再在空中画出一个平面，这样就把蓝点和红点分开啦，如 19.2 右图所示。

事实上，真实神经网络所做的事情与之类似。如此一来，一个线性分类器（解码器）所能分类的种类大大增加，也就是说，我们得到了比先前强很多的模式识别能力。此处，高维即高能。

那么，神经编码的高维度是如何得到的呢？仅神经元的个数多是没有用的。假设我们有数量庞大的 *N* 个神经元，而每个神经元的放电率只与 *K* 个关键特征线性相关，那么我们最后表征的维度只会等于问题本身的维度，*N* 个神经元毫无作用（多出的神经元都是前 *K* 个神经元的线性组合）。如果要突破这点，就必须要有与 *K* 个特征非线性相关的神经元，这里我们称之为非线性混合型神经元，这类神经元的表征十分复杂，而其原理正类似于 SVM 中包含非线性项的核函数。有了这些非线性的神经元，神经编码的维度才可以突破任务特征的维度。

如图 19.3 所示，神经元 1 和 2 分别只对特征 *a* 和 *b* 敏感，神经元 3 对特征 *a* 和 *b*

的线性混合敏感，而神经元 4 对特征的非线性混合敏感。最终，只有神经元 1、2、4 的组合才能使得神经编码维度升高。

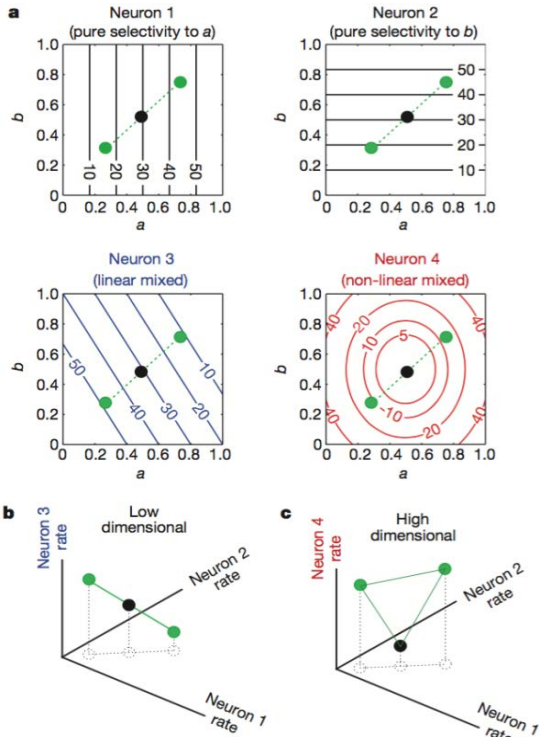


图 19.3

这种编码的官方叫法是混合编码（Mixed Selectivity），在人们没有发现这种编码原理时，我们觉得这是不可理解的，因为神经网络对某种信号的响应显得乱糟糟的。在周边神经系统里，神经元的作用如同传感器，对信号的不同特征进行提取和模式识别。每个神经细胞的功能都是相当特定的，比如视网膜的 Rods 和 Cones 就负责接收光子，而之后由 Gangelion Cell 继续进行编码，每个神经元就好像是一个个被专门训练的哨兵。而在高级脑区，这种清晰的分工难以见到，我们发现同一个神经元可能对各种特征敏感，而且这种敏感还不是线性的。它们更像是对各种任务都想掺和一下的“万金油”，这种很难找到线性可分的专业分工的现象，在我们对机器学习中的 SVM 方法做了对比后才清晰起来。原来，这正是对原有的信号做了非线性变换（如果 $x1$ 是一个特征，$x2$ 是一个特征，那么这种神经元可能就是 $x1^2+x2^2$），而使得神经编码

的维度得以高于信号特征空间维度的办法。

大自然的每个细节都内藏玄机，大量冗余和混合编码这看似不专业的做法，看似混乱的信号，最终得到了更好的计算能力。有了这个原理之后，我们就可以轻易地处理这样一些任务，如图 19.4 所示。

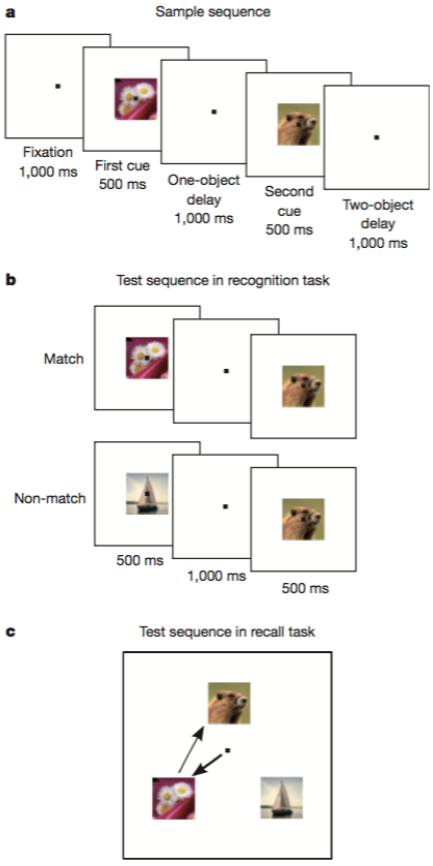


图 19.4

在这个任务中，猴子首先被训练分辨一个图像是否和之前的相同（Recognition），之后被训练判断两个不同图像出现的顺序（Recall）。猴子要想完成这个任务，就要能够对任务的不同侧面进行编码，比如任务类型（Recall or Recognition）、图片种类等，而这正是绝佳的测试是否有混合非线性编码机制存在的实验。实验中证实了大量神经元确实对混合特征敏感，而且存在非线性（比如说同样是对花朵进行编码，神经

元放电强度取决于任务是 Recall 还是 Recognition，特征之间不独立）。混合编码使得神经编码具有高维表征的特性，从而让这些包含了多个侧面任务的解码和处理得心应手。

因此，在设计神经网络时如果引入一些非线性的单元会大大提高模式识别能力。SVM 恰好应用了这点，来处理非线性的分类问题。而计算神经科学与机器学习，犹如一枚硬币的两面。

我们研究脑区的功能，先要用机器学习的方法处理数据，比如用 PCA 找到问题的关键维度，之后再用机器学习模式识别的思维理解神经编码和解码，最终如果得到了一些新的灵感，我们又可以改进机器学习的方法。对于大脑还是机器学习算法，最重要的都是得到信息最恰当的表征方法。有了好的表征后，就做什么都容易了。这正是机器学习从线性逻辑回归到支持向量机，再到深度学习的一步步进化过程，或许这也是大脑得以进化，我们得以对世界具有越来越高的把控能力的过程。也许进化的本来目的就是更清楚地分清谁是老虎谁是羊，谁可以吃谁，而在此过程中，却发展出对世界本身步步深入的理解，以及对理解本身的热爱。

20

机器学习是如何巧妙理解我们大脑的工作原理的

（阅读难度★★）

机器学习和复杂系统正在诸多领域改变产业界和学术界的传统方法，其中一个极好的例子正是给机器学习以巨大启发的神经科学本身。在此我们用一个鲜活的例子来展示这个方法在该领域的展开。

本章的开头想问大家一个问题，如果给一只猫建立一个模型，那么最好的方法是什么？这个问题很深刻，大家可以边读下文边寻找答案。

神经科学，是一个高度依赖数据的生物学分支，因为大脑本来就是用来计算的装置，要想研究它，首先要研究它的输入和输出，然后把它和动物的相应行为联系起来。

这个故事的第一部分与复杂网络有关，因为神经系统本质上属于复杂网络的一种。我们希望通过在电脑上建立一个和大脑神经网络类似的复杂网络，来理解神经细胞数据到行为间的联系。这种努力在 20 世纪 80 年代至 90 年代十分盛行，比如最早出现的 hopefield 网络解释记忆现象等。

下面给出一个典型的例子。

科学家发现猴子可以经过训练对不同频率的两个声音进行比较(出错应有惩罚), 这是一个典型的测量短期记忆能力的实验。如果猴子需要比较两个声音，就要把第一个声音的信号放在脑子里，然后和第二个声音进行比对。能够综合不同时间的信息进行决策说对生物生存至关重要，而此实验即为其基础。

以往神经科学的研究方法可能只是描述这个行为，然后想办法在猴子的脑子里放入电极，测量相应的神经信号是什么，如图 20.1 所示。

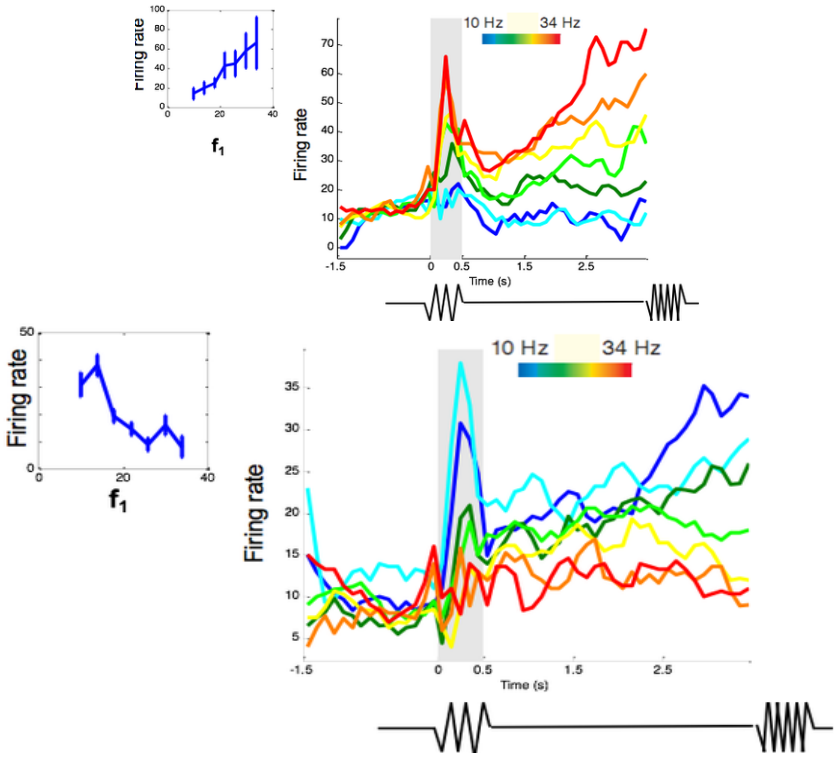


图 20.1

例如，我们可以看到图 20.1 中测量的神经元放电信号，这个神经元放电的频率随着声音的频率上升而上升，因此我们就可以根据它臆想一个模型出来。图 20.2 测量的细胞则相反。那么能不能根据这两个图设计一个模型来解释猴子的行为呢?

可以！最简单的物理模型就可以解释这个现象。在图 20.2 中，左图用山谷里的小球来描述这个模型（算法），这是把整个问题放在极低的维度上以方便理解，整个外界输入（声音）可以看成是在猴子神经系统上的一个外力，外力可以看作某种推动力，让猴子的神经元状态发生定向改变，犹如小球（猴子神经系统的状态）在一个被外力塑造的山谷里趋向谷底（出现概率最大的状态，对应物理里能量最低的点），这个谷底的位置就是对第一次信号的记录（记忆），当外力（$f1$）消失，山谷的地形发生了变化，但外力并不马上改变。当第二信号（$f2$）来到的时候，地形再次发生改变，此次的信号引起山谷的隆起，导致小球滚向新的谷底（左右各一个），而小球最终达到的位置就不仅与这次的外力有关，还与之前外力引起的位置高度相关（综合历史信息进行决策），这无形中就实现了对两个信号进行比较（此处即最简单的图灵机）。那么如何利用我们刚刚说到的两种神经元（和外界信号正比或反比）来实现这个功能回路呢？请看图 20.2 的右图，我们甚至可以画出一个电路图来解释这个原理。正号代表正比神经元，负号代表反比神经元，E 代表一个随时间变化的控制信号，S2 根据 E 改变电路连接，因而可以设计一个带有记忆功能的减法器来实现它。

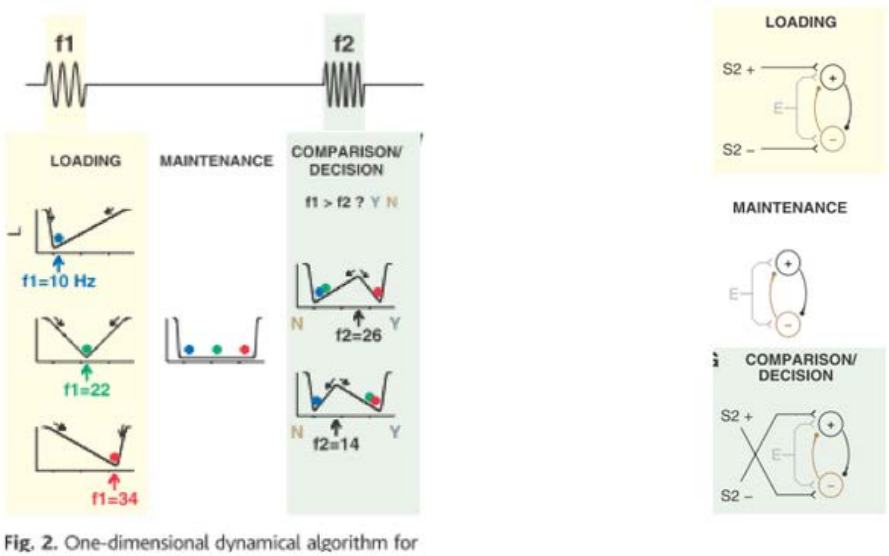


Fig. 2. One-dimensional dynamical algorithm for

图 20.2

这个方法本质上就是一种类比，很多真实的神经科学家把这个方法戏称为 toy

model，而对其不屑一顾。因为无论你电脑里的程序行为多么像大脑，其实与真实都相差万里，而且能够实现某种功能的算法也有很多，凭什么说大脑就是按你的臆想工作的？用这种方法的计算神经科学家经常纠结于自己的模型需要多大程度仿真这个问题上，就像有些人说的，要想给一只猫建立模型，最好的方法就是找一只猫来。

当然有模型还是比没有模型要好很多，毕竟它点亮了生物世界和数学世界的联系。使得一个我们可能理解的数学体系得以建立在烦琐摸不到体系的生物体系之上，让我们能够通过改变参数空间的方法与之玩耍。

然而模型开始发挥威力的真正时刻是机器学习的介入，传统的复杂系统方法从此得到革新。为什么呢？

因为通过机器学习，我们拥有了一个强大且精确的从数据反推模型的方法，而非之前模糊的类比。机器学习的思路和之前的根本不同在于，我们不再醉心于设计一个与真实系统相同的模型，而是先用与真实模型原理大致相似的标准化模型，去学习真实的输入和输出。此处的思维即你不再追求画出一个活灵活现的猫，而是先做一个四不像的东西，让它去学习和猫一模一样的行为，当这种行为真实到不可区分时，你可以认为它就是那只猫，如图 20.3 所示。

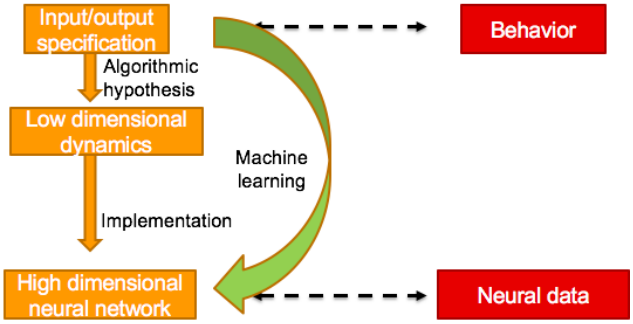


图 20.3

一旦机器学习开始介入，模型就被赋予了预测性，而被赋予预测性的模型，可以看作在输入输出层面与真实系统不可区分，从而解决了模型复杂度不好的设定。机器学习的模型比之前的模型更好的点在于多了 Cross Validation 的部分，用真实数据得到的模型，不仅要在已知输入输出的情况下做到像真猫一样，还要在未知输入的情况

下像一只猫，也就是说，它真的要有学习能力，能够像真猫一样不仅可以捉到屋里墙角的老鼠，还可以捉到田间地头的老鼠。

用这个方法，我们的研究框架发生了变化。首先，我们要寻找一组具体的输入输出作为研究起点。比如要研究视觉区域，就要找到一组猫和狗的照片，然后，输出正确的分类，再进行测试。

与复杂系统模型方法不同的是，机器学习的方法是一个黑箱操作的思路，我们首先做一个标准化设备，然后把大量输入送到这个标准化设备里，然后改变参数得到一组我们想要的输出。黑箱的好处是训练得到参数的方法已知，因为这往往是一个十分艰难的任务，与传统复杂系统模型只追求定性描述的需求不同。

如果上面的问题用机器学习的方法来解决，则如图 20.4 所示。

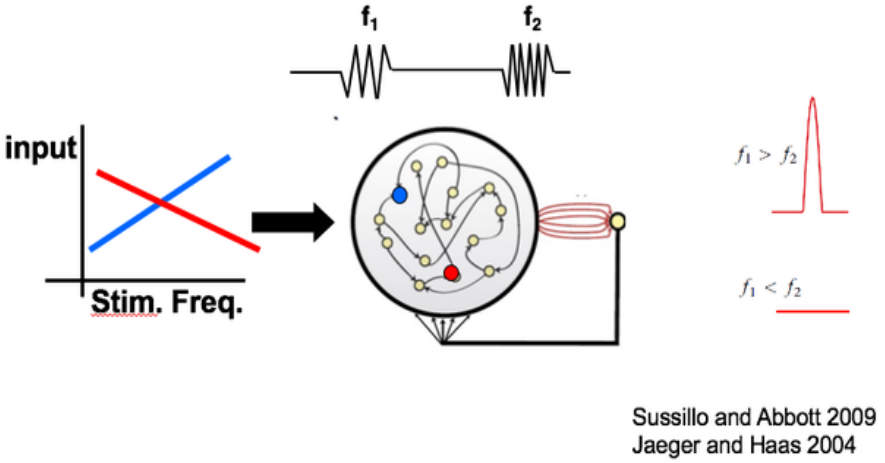


图 20.4

此处的主角是一种标准化的神经网络 RNN（循环神经网络），我们给定在各种情况下输入到输出的映射关系，然后在有监督学习的框架下对网络进行训练，即调整网络连接权重达到模拟这个输入和输出的过程。而此时我们不需要设计，两种与实验相符的神经元就会自发浮现出来。也就是说，机器学习出来的猫完全具备了猫的所有功能，甚至不用我们过度描摹，就长得也有些像猫了（功能和形态的对应）。

现在我们有了一只会在各种场合抓老鼠的猫，但我们不需要一只机器猫，而是要

通过机器猫研究猫的行为，懂得猫是如何通过底层的元件实现功能的。电子猫比真猫要听话得多，可以让我们任意解剖，改变参数。如何做呢？再次回到复杂系统里的动力学分析，如图 20.5 所示。

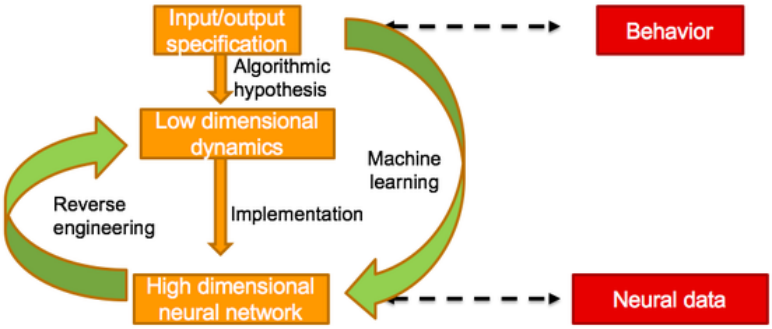


图 20.5

此处新增的箭头是从高维神经网络到低维动力学系统的。还记得我们之前说的那个山坡小球的玩具模型吗？它的机理很完美，然而可惜的是大家并不信服，此处我们有了再次证明的机会，取出训练好的能够与真实系统不可区分的 RNN，然后用它进行“逆向工程（Reverse Engineering）”，从中取出一个低维度系统，看它的工作机理是不是符合我们的猜想。至此，我们用到了一个非常抽象的方法，用真实数据回答了一个几乎不可能验证的假设。

这里其实还有一个隐含的信仰，被训练过的黑箱被认为学习了生物系统的本质。你的 RNN 做任务做得再好毕竟和真实还是有区别的，但此处我们已经达到了目前可以做到的极限。这里涉及一个十分根本的问题，就是机器学习的核心，即通过数据学习得到真实系统的数学表征，而这种表征有多大程度接近真实，依然是难以量化和描述清楚的。

通过这种高维 RNN 到低维动力学系统的映射，我们就得到了一个解释整个系统运作的更加简单的示意图。通过学习，一个高维混沌系统开始出现定点（Fix Point），定点含有系统输入输出的重要信息。此处是一个简单的二元分类器，这个分类器工作的原理正是一个动力学里的鞍点。对于 $f1>f2$ 和 $f1<f2$ 两种情况，我们得到一个分叉行为（Bifurcation），一个会趋于是回答，另一个会趋于否回答。某种角度，这也是机器学习里连续到离散的分类问题的一个动力学解答。其中的物理含义与最初我们推

倒的那个猜想玩具模型是一致的，如图 20.6 所示。

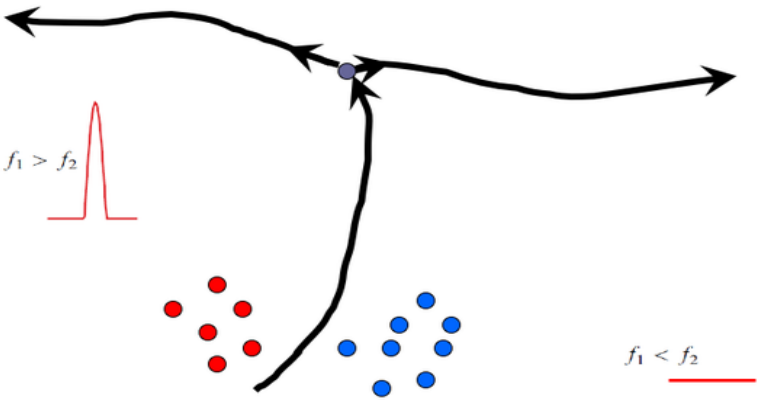


图 20.6

从中我们看到了机器学习和复杂系统你中有我我中有你的关系。首先，机器学习是一种算法，而算法总要有一个物理实现（复杂系统）。我们先有数据组成的表象世界，再有算法组成的符号世界，最后是抽象的真正解释世界机理的物理世界。这三者之间，表象世界通过机器学习进入符号世界，而符号世界又能与物理世界巧妙地相通。机器学习是桥梁，复杂系统是灵魂，而没有桥梁，灵魂就是空洞的。

我们可以进一步追问机器学习方法如何工作，人脑如何工作，此处真正相通的地方到底是什么？注意，我们反复在说的低维到高维与高维到低维的问题，现实生活中的问题往往是高维到低维的映射，比如信号（高维）——决策（往往二维）过程，能够在高维空间里找到低维嵌入，往往就代表神经网络建立了真实世界的模型，正是因为这些模型，我们才具有了“举一反三”和“泛化”的能力。更深刻的，就不是本章所涉及的了。

21 大脑经济学

（阅读难度★）

我们每天接收无数外界输入（感觉、信号），但是只有被我们大脑阐述为有序的信号才叫信息，其余则是噪声。为什么对于某些信号常认作是信息，而有些却认作是噪声？因为我们的大脑结构天生能够对某一类的外界信号建立有序，抽取信息，而对其他信号麻木。

科学家沉迷一套缜密的逻辑，因为他们的大脑对因果关系特别敏感，例如贝多芬把音符组合成美丽的旋律。若是在大脑天生容易产生信息的地方进行造化，那么你就可以成为天才。而如果硬要在那些你的大脑天生比较容易产生噪声的地方建立秩序，则反抗的将是你自己。

生物在世界上存在的第一个矛盾是能量的稀缺，尤其是那些处于低熵状态能够被利用的能量就更少。这也是所有生物竞争和进化的动因。能量有限，欲望无穷。大脑以 2%的重量消耗人体 20%的能量，因此能量稀缺是大脑面临的第一难题。哪里有稀缺性，哪里就有经济学。

Lesson 1：人性贪婪，贪婪的原因是稀缺，因为稀缺，所以只有有竞争动力的个体才允许存在。

Lesson 2：人脑的经济学。经济学是市场上无序的个体在自发组织下形成有序的市场，并通过价格杠杆实现资源的优化配置，用最少的资源做最多的事情。那么大脑的经济学是什么？用有限的能量，产生最多的信息，这就是大脑应对稀缺性的对策。

Lesson3：市场通过给买卖的人排出优先次序，让最想付出代价取得货品的人优先赢得交易。而大脑通过情绪给经过它的众多任务（信息源）排出优先级，分配能量。大脑喜欢给那些它擅长阐述的东西赋予能量，是大自然优化配制的表现。有一个词叫舒适区，就是指这一类任务。

读懂情绪所包含的信息

我们常说冲动是魔鬼，情绪有如此大的力量，是因为我们需要它。情绪来自我们亿万年形成的本能，如同一头威力巨大的原始野兽，让我们愤怒的时候攻击，喜欢的时候坠入爱河，成为一个合格的生物人。

情绪是大脑调节能量流动的方式，正如价格是市场调节物资分配的方式。你所做的不同的事情，有的喜欢，有的不喜欢，其实你的喜欢与不喜欢就是大脑在估价。大脑要把稀缺的能量分配给不同的任务，这个时候它要对每一件事进行价格评估，看看哪一件的效用最高，就给它传送能量。这里不得不说大名鼎鼎的多巴胺，如图 21.1 所示。估值的任务就是靠多巴胺提供的，被多巴胺评级高的事物，大脑就给它更多能量，从而提高大脑的信息转化率。

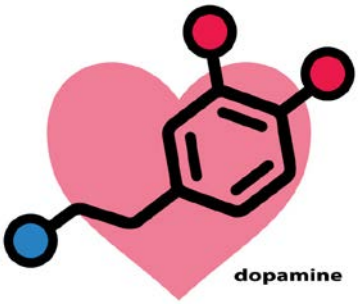


图 21.1

每个人的估值体系不同，这决定了一个人的根本性质和所适合的生命结构。一个典型的例子是内向和外向。内向的人也可以很善谈，但是却不喜欢和别人应酬，这点

情绪已经做好了估值。读懂情绪，就是理解大脑自身的估值体系。

但是，人是否可以违背情绪而行事呢？当然，因为我们有自我意志。这是人类作为万物灵长最有利的武器，我们的自我意志可以计算长远的利益而抵触我们的本能。但是，当你抵触情绪执行任务时，消耗的能量要比顺着情绪来大得多，因为你需要一个高级的中央控制系统来抵消情绪的趋势，这本身就是耗能的，再加上违逆情绪的任务一般都是信息转换率很低的，因此情绪会通过痛苦来告诉你它的反抗。

压抑大脑自发的估值体系而做事情的经历相信大家都有。

比如说，王小五十分痛苦地背书，为了记住热力学第二定律，他费了九牛二虎之力，最终忍受不住去看电影了。为什么？白话说对物理没有爱。但是爱这类的词只是表象而不包含太多信息。其实是他的大脑在物理这个渠道上能量流动不畅，费了九牛二虎之力，刚吃的三斤馒头的能量都用上了，那些书上的文字依然只是噪声，即形成了非常少的有序性。这个时候大脑发出警告信号——痛苦，我不适合干这个，换个能量渠道吧，恰好旁边有电影，大脑里瞬时发出滚滚电流。

当你感到疲劳时，就是情绪的负反馈，大脑让你离开现有的工作模式，这个模式的效率已经太低。而你感到开心，则是一个正反馈，即加大能量在这个渠道上的投放剂量。

因此，寻求快乐的本质是找到那个开启大脑正反馈回路的钥匙，而懂得读取情绪所包含的信息，就是第一步。

顺着内心的快乐做事，从根本上看是节能高效的措施，就像根据市场价格买东西。记得有一本书说人要是长期做一件事情超过多少小时就是专家，这在理论上或许可行，但却不容易实践，因为做一个你不喜欢的事情就是消耗过多的能量而无功，最终进入痛苦的恶性循环直到无法持续。

尼采在他的名著《悲剧的诞生里》用酒神和日神比喻情绪和理性。酒神代表的情绪是原始丰饶，是艺术灵感，即强健的生命力的源泉。而日神代表的理性则象征着文明、秩序与逻辑，它会导致原始丰饶的丧失。酒神虽然美好，但日神却是必需的。

光靠情绪行吗？

这归根结底是个调节短期和长期快乐的问题。大脑的情绪调节是一套来自石器时代的技术（进化的滞后性），那个时候除了打猎、交配和繁殖，我们没有其他的任务，而且，人也活不长。那个时期的技术放在信息时代，至少不完全可靠，因为，它永远只看眼下，不会平衡短期和长期的快乐。

完全跟着情绪走往往等于无原则的寻环作乐，想想那些我们在百无聊赖的时候最喜欢做的事，是不是都和追逐打闹有关，它反映了一种原始人打猎和繁殖的需求。但是这一类得来的快感没有可持续性，还会抑制你发现那些具有持久效力的方式。因此长时间看它们并不经济（竭泽而渔的原始刀耕火种经济）。

为了长期的繁荣，我们需要法律来规范市场，我们需要银行来调节长时间的流动性。情绪需要和理性自律配合，标志性思维是以情绪为信息指标，但不完全是决策标准。

对于大多数有格调的事情我们并不能在一开始感到快乐或者自发的开始，就好像物理中所说的势垒，我们要先投入一个启动能量，才能突破它，达到极乐的香格里拉。如果不投入能量，我们自发开始做这件事的可能性是零。如同一些开始时没被投资者良好估值但长远来看是黑马的项目，它们需要一定的启动资本才能显示价值。投资者就是分配能量的大脑，而大脑天生吝啬，不喜欢投启动资本，结果埋没了这些项目。银行的贷款就是投资启动资本，如果他们真的是潜力股，后期就会脱颖而出。

大多数需要长期培养的兴趣都属于这一类，比如钢琴、绘画、数学、物理等。

这种为跳出舒适区，开启新的大脑正反馈通路的成本叫低能势垒。无法突破低能势垒，是很多人无法达到自身最佳状态的原因，这时人就被困在一种简单无脑的惯性状态里。

如果把大脑可能执行的任务画成一张图，并且把达到每个任务的实现难度标记为需要攀登的高度，那么我们就可以做出一张大脑认知的能量地图（Energy Landscape）。眼下的舒适区就是这张地图中你目前所处的谷底。前面说过，大脑偏好节能，因此，你在谷底也会感到特别舒服。然而在这个位置上，大脑真正的潜能却处于闲置状态，如图 21.2 所示。

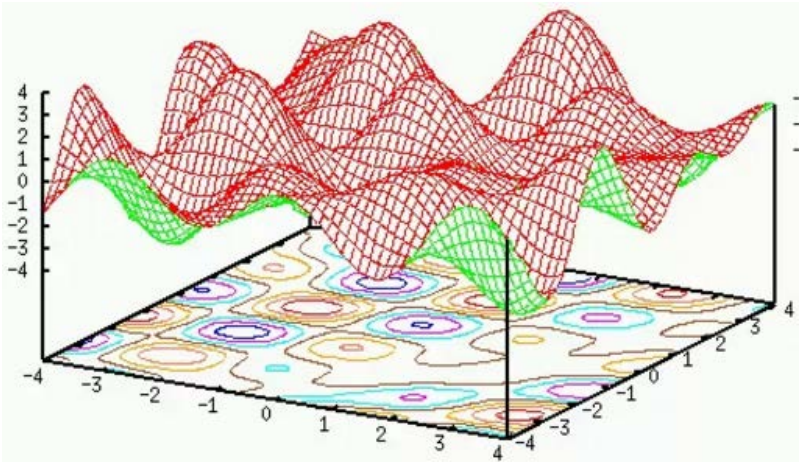


图 21.2

舒适区可能与真正的那个可以吸收整个生命能量的最优点相差甚远，但是却极难跳出。一个人所能达到的人生建构，取决于他对自己完整的能量地图的认知，以及他是否能够努力找到那个真正的最优点。

心理有三区：舒适区、挑战区和恐惧区。挑战区是自我调整的关键。

克服势垒的方法是 Hebbian Law。还是之前说的，情绪的力量很大，如果你一开始想要跳出你的舒适区，将会得到情绪的玩命儿反抗，而再次掉入谷底。因此，你要做的不是马上跳出，而是用习惯修正你的能量地图，减小低能势垒。做法就是根据一个最基本的神经学原理：你的神经活动也可以改变你的神经元之间的连接，经常在一起放电的神经元会加强连接。

还是回到大脑要做的两个基本任务：感知和决策行动。大脑的决策取决于神经元起初的连接，又称条件反射。这些决策相比通过自我意志引导的复杂运算的决策，耗能要小得多。为什么会有舒适区，是因为在那里你只需要依靠预先编码好的条件反射就够了。要想离开舒适区，就需要弱化这一区域有关的神经联络，或加强新的区域的神经联络，这样就可以减小低能势垒。

Hebbian Law 的生效往往需要一段时间，这也就是我们所说的习惯养成是长期的过程，需要强大的理性自律。但归根结底，并不是你可以任意地改变这些联络。合理的人生目标就像一颗种子，你的大脑是这颗种子的沃土，后期才能发芽长大。或者说，

当你离开了舒适区，你需要确实寻着你心智能量地图的最优点去，否则只能在痛苦中夭折。这一点只有一边行动，一边观察自己的情绪，才能搞清自己是否在正确的方向。那些长久让你抑郁的东西是不能通过习惯变成你的长处的。

图 21.3 所示的是放电导致的神经突触加强，即两个神经细胞之间的联络加强，更加容易一起放电，这就是神经可塑性与 Heibbian Law。

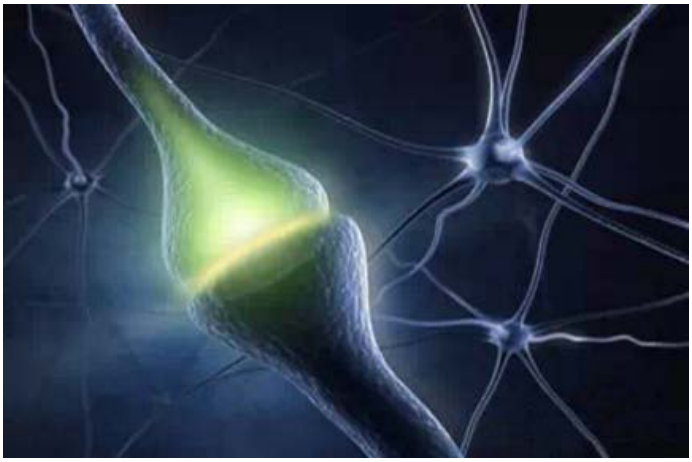


图 21.3

用一句经典的台词结尾：Your brain is programmable, if you don't program for it, others will（你的大脑是可编程的，如果你不自己编，会有人帮你编好——那人极有可能是撒旦的使者）。

在舒适区和挑战区的边缘寻找突破点，解放情绪的力量。

最后回到主题，如何开启大脑的 high 模型，汇集发散的欲望流于一序？刚才说过，自我意志控制的行为具有高耗能的特点，物理里说叫克服阻力做功，这里则是克服情绪做功，但不能持久。如果能够反之，把情绪的箭头反向，那么阻力就是动力，这件事，早有先人通过心流的概念描述过了。

对于心流的定义，维其百科是这么说的：

“flow,‘also known as zone, is the mental state of operation in which a person performing an activity is fully immersed in a feeling of energized focus, full involvement,

and enjoyment in the process of the activity. In essence, flow is characterized by complete absorption in what one does.'"

这里说到，人进入心流状态后会进入全神贯注的状态，感到一种内生的喜悦，自我耗能被所做的事情吸收掉了。心流的一个根本特征是 Intrinsically Rewarding，意思是说你所做的事情本身使你得到补偿，而不是因为它能实现你要的外部价值，如赚钱。

在心流状态里，情绪起到润滑剂，也就是正反馈的作用，连续不断的多巴胺产出，让你欲罢不能，以致忘我，连失恋这样的事都可以忘记。想必很多人在游戏中都经历过。

心流产生的条件是什么？是一个感知和行为的正反馈。一个人执行任何任务，都可以分解为一系列小的收集信息和决策的过程。你感受到的信息立刻在大脑里激起下一刻要出现的画面，并因势利导做出行动，因此得到理想的结果，再通过感知反馈回来，刺激多巴胺的分泌，给这一任务更多的能量，鼓励进一步行动，周而复始，直到你调动了所有的大脑能量。

反之，如果你不能阐述收集到的信息，而做出错误的预测，导致错误的行为，那么你大脑分泌的化学物质就让你不爽，不要继续下去，从而精力分散。

极限运动中的人一般都处于心流状态。比如冲浪，优秀的冲浪手盯住即将到来的浪花，脑中已经想到了它下一刻的走势，并顺势做出一个动作，当他的决定让他站立在浪潮之巅时，他内心的喜悦也达到顶峰，近乎一种忘乎所以、天人合一的状态。心流中的人处在一种感知、行为和反馈完美衔接的状态。

所以心流产生的关键因素有三个。

一、心流容易产生在舒适区和挑战区交界的地方，如果大脑越级挑战，则多半会被打回去。而如果仅在舒适区，一时的快感最终一定会被无聊和空虚取代。

二、是否及时得到反馈，因为多巴胺的产生是你决策得到的结果与预期进行比较得来的，如果你的行为不能很快产生结果，那么正反馈也很难被激活，当然这点因人而异。

三、最核心的还是你的大脑对特定信号的阐述能力，你可以极为有效地在信源里

提取信息而不是噪声。心流属于作曲中的贝多芬和投篮中的乔丹。

心流中产生的大量多巴胺同时又和 Hebbian 定律说的神经系统的可塑性相关。如果你在一个任务里启动了心流，那么你多半会对它“上瘾”，并因此加固心流。这就是一个人在某个领域形成精深造化的基础。他所能达到的造诣有多精湛，跟他处在心流状态的时间成正比。

心流需要突破低能势垒，以及寻找自己大脑的“敏感区”。最后说一下如何找到大脑的敏感区。

你的大脑到底对哪一类信号敏感？观察情绪给你的反馈，是语言（故事）、数学（桥牌？魔方？）还是运动？这只能从你的生活经验中得来。

但是这种方法获取的信息往往很少，因为很多人可能直接掉入打游戏一类。这些事情被选择是因为太容易了，势垒为零。它们能被选中仅仅是因为要求的心理能量较低，而非真正的最优解。如果你被这一类事情迷惑了，很可能就是掉进了低能陷阱，即你以为是跟着你的心走，实际上是躺在心灵的谷底睡大觉。

所以有了第二条路：在一些生活经验的基础上对此进行剖析，把复杂事物分解为基本元素。那些复杂的任务包含的基本思维过程是不是你喜欢的呢？

笔者的假设是，任何一个人生的范式，都可以被分解成某种节奏性的信息。比如你所做的事情给你反馈的频率。

比如一些人是无法做科研的，因为它是一个反馈节奏极其缓慢的事情，你的大部分时间所从事的是单调的基本试验，而且不确定度极高，你几乎不知道结果的好坏。但是科研的好处是它能最大地调动你的创造性潜能，而且不停地进行因果性思维。如果你对因果敏感，容易满足于自己内生的思想，那么科研很可能是你的范式。

商业的事情就是比较多互动，节奏比较快，见效也快，但风险大，适合风险偏好者。

还有一些技术工种，风险和不确定度低，节奏稳定。

从上述几个例子我们可以看到，其实生活的范式是可以分解成元素的，如风险高低、节奏快慢，所动用的思维是形象的还是逻辑的，与人互动的程度等。而其中每一

个元素是否符合你，都可以通过更简单的任务测试得到，即生活细节里你自发的趋势。你是比较求稳还是求变？你是比较好进攻还是比较被动？你思考是比较喜欢图像、数字还是语言？这些自发的趋势会告诉你比较擅长发掘哪些事情的有效信息——人脑天生的结构。

寻找到那个恰当的生命之序，并用心流把无序发散的心理能量联合起来，你就解决了人生最大的命题。其他的问题，如人生的目的、意义等都成了钻牛角尖的废话。你只需要进入你的范式，人生必然精彩。

有趣的推论

（1）做事的顺序，即根据精力大小排列不同困难程度的事情。那些较为困难的事情，只有在精力最旺盛的时候才比较容易激发能量流。如果用最好的精力做了一般的事情，那么消耗精力就是浪费。

（2）心理疾病。很多心理疾病，如抑郁症，就和前面讲的被困在某些心理过程的低能级区域有关，由于受困者缺乏突破势垒所需要的启动能量及方向，而身陷其中。由此可以引出心理疾病治疗的一些方法。

（3）知行合一。生命过程是感知及行动的一系列连锁反应。感知是大脑中建立的世界的模型，这个模型将指导你行动，行动过程中你会得到反馈，看你的模型是否正确。人们在知和行完美结合的状态中感受到的快乐是最大的。

（4）低能陷阱。人很多时候的自发选择只是因为那些事情最容易做到，事实上，我们也的确在疲惫或者能量不足时，抵抗熵增的能力减少，而易于处在充满不确定性的无序状态。这给我们反过来利用低能陷阱提供了可能。

比如说追求一个比较难追的女孩，可以反复向她提出吃饭的请求，头几次可能都是拒绝的，但是拒绝别人是个比较耗能量的事情。等到她能量下降到一定程度，就有可能掉入低能陷阱。这时候她的回答的不确定性增加，因而说 yes 的可能增大。这也是为什么在她们疲劳或脆弱的时候提出一些请求比较容易。

22 人工智能 vs 人类智能

（阅读难度★★）

深度网络代表的这波人工智能风潮代表的是一种局部的演进，还是打开了未来的一扇大门？

AI 离真正的智能很遥远是因为它不能创造吗？是因为它只擅长形式逻辑？是因为它没有自我意识？

其实，我们连生物的智能都不了解。现在，我们回顾一下神经科学和 AI 的历史，从脑和智能的演进来看这个所谓的人工智能有多聪明。

要谈这个问题，我们就需要从智能说起。智能其实分为三个层次，对应丹内特对意识进化的三个分级：达尔文式造物、斯金纳式造物和波普尔式造物。

智能的第一个层次是进化（达尔文式），并不是自我意识这些高端的东西。智能的起点是学习，学习即对环境变化做出相应对策。整个生物进化过程，就是学习的过程。为什么呢？一个小小的细菌，也可以对环境做出趋利避害的反应，并且通过基因突变的方法有点盲目地适应环境，这其实就是用遍历法来选择针对环境变化的最佳生存策略，然后通过遗传以及下一代继续试错，将某种策略强化。AlphaGo 的策略网络也是类似的道理，通过对可选策略集合的分析进行局部最优调整。细菌和十亿年的恶

劣环境下棋，把对哪些化学物质该如何转化这个信息深深地埋藏在了它的 DNA 里，因此才有如今从极地到大漠如此伟大的适应性。大自然的这种学习方式可以看作智能 1.0 版，缺陷是速度慢和读取数据量小。我们人类模仿进化的过程创造了进化选择算法，是人工智能一个非常基础的部分。

智能 1.0 版进化算法一一遍历所有可能并加以选择，最终得到最适合环境的策略，被动但是在历史长河里极为有效。

达尔文智能主宰了漫长的生物进化史的大部分岁月，使得单细胞生物逐步具有越来越复杂的性能。

AlphaGo 的策略网络（蒙特卡洛树搜索）——细菌级别的智慧，20 年前的深蓝就使用它战胜了人类国际象棋大师。

下面介绍第二阶段的智能——斯金纳（伟大的行为心理学家）式造物。

斯金纳式造物说的是生物自己能够自主地去学习而非被动的靠基因变异适应环境。这项伟大的创举背后就是大名鼎鼎的神经网络，生物进化几十亿年的历史都是这种被动地适应环境，直到神经网络的出现，一切才悄悄发生变化。

神经网络的作用简单来看，就是一个分类器，它把外界刺激分成好的和坏的，并且趋利避害。拥有这个分类器，动物终于可以在自然环境面前主动做决策，并且趋利避害。把狮子放在要躲避的那一堆，异性放在要接近的那一堆，这对于动物的生存意义之重大不言而喻。这个分类器最开始是储存一些先天的条件反射，比如婴儿见到乳头就要吸。而后来就出现了后天习得的条件反射，比如著名的巴甫洛夫的狗，听到铃声就会分泌口水。没错，后天形成的条件反射就是学习的 2.0 版。

然而生物神经网络是如何实现这一调整的，却一直是个迷，直到 1940 年 Hebb 提出神经科学的牛顿定律：Hebbian Learning Rule，人们才开始了解神经网络是如何实现这一步骤的。Hebb 说，组成神经网络的神经元通过不停地调整突触连接来改变对外界刺激的反应，这个变化法则就是同时放电的神经元连接加强（细节来看还与放电的顺序有关）。这就使得被一起激活的神经元形成一个基团，比如狗听到铃声以后被喂食，起初铃声这个刺激对狗可能没有任何反应，而之后就被划分到用餐那一类，从而形成对铃声的条件反射。AlphaGo 深度学习的基本原件人工神经网络也是一个人

为敲定的分类器，用于做决策。

人工神经网络的训练过程同样借鉴了生物神经网络的学习过程，根据反馈调整神经元之间连接的权重关系，来实现对外界信号分类方法的改变，由此调整决策（Reinforcement Learning 强化学习）。

生物形成新的条件反射，可以理解为一个决策的生成，用二进制表述后成为分类问题，如图 22.1 所示。

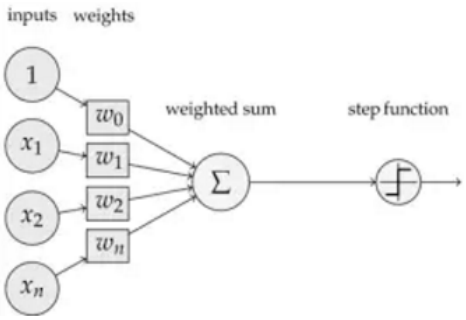


图 22.1

分类问题可以由神经网络解决。图 22.1 所示的神经网络，就是一个基本的卷积网络，把输入的值乘以一定权重再加在一起，然后通过一个非线性的阶梯函数，转化为 0（有害）或 1（有利）的输出，即决策过程。

数学上叫调参：改变 w 的数值即可改变分类的方法（界限）。随着数据量的加大，分类越来越准确。

神经网络的分类功能，把输入的信息（环境变化）分为有利的和有害的进行决策，环境的变化越复杂，越能体现神经网络可以任意通过改变连接强度来调整决策“界面”的优势，而不需要用进化的方法来适应环境（学习的重要性）。当然，这个学习过程需要大量数据的训练。

结论：一个单层神经网络完全可以娴熟地应用斯金纳式造物。

从第二种智能方式我们可以看到，生物智能的方式是如何启发了人工智能的。

智能的最高级形式是波普尔式造物，对外部世界进行表征，形成认知、信念和预

期，对应了神经网络的更高级功能。

如果仔细思考，会发现这些很多包含在 AlphaGo 使用的深度网络里。深度网络最擅长的是对事物进行抽象（深度学习），在最靠近输入的层次上，每个细胞就如同数码相机 CCD 上的像素，之后的每一层次都比上级网络的感受野要大，而最终得到的效果是最深层的神经元直接处理和图像的全貌相关的特征，比如照片上的人是谁。如果换到其他地方，就是从抽象或全局特征进行决策。

这个结构极像人类社会的结构，越是高层，越能把握和掌控全局。深度网络上的“抽象概念”这个认知武器，使得 AlphaGo 有对全盘棋的趋势进行判断的能力，可以迅速舍弃一些错误的方向，减少搜索的深度，即价值网络。其实人脑所使用的算法和 AlphaGo 差距没有那么大，记得前些年有一篇著名的文章说人类发现在高级脑区表现抽象概念，如人名的细胞，这是符合这种深度网络逐层抽象的概念的。高级脑区正是对应人脑深层网络的最底层。

AlphaGo 能够战胜代表棋牌巅峰智慧的围棋冠军这件事最大的意义，也在于深度网络赋予了 AI 自主判断局势和形成策略，而不是之前的仅靠人为给定的策略遍历所有可能。或者说深度网络打开了波普尔式造物的大门。

当然，深度网络算法只是提取了生物神经网络的一个主要特征，而几亿年进化结晶的人脑，由于计算机能够提取并用于学习的数据量巨大，使得它能够在学习了人脑的一个雕虫小技之后通过迭代学习迅速在某个特定任务上超越人类。

就上面这个简单的历史陈述我们发现，AI 说到底是一种仿生，但是这种仿生无疑会改变我们生活的方方面面。AlphaGo 的智慧是结合了古老的细菌智慧（策略网络）+高级哺乳动物的智慧（价值网络），可谓仿生物的智慧杰作。

一些常见问题如下。

AI 是否能取代人类大量劳动

AI 一定会逐步取代简单的人类劳动，但是也会增加新的劳动出来，比如 AI 设计、AI 纠错，以及如何利用 AI 做出以往实践不能的事。AI 将使得人脑从简单劳动中解放，可以爆发出不可预计的新产业。

AI 没有意识，无法预测，没有创造力，AI 几乎永远无法与人类智能相提并论

人类容易犯的错误之一，就是用一些自己也无法严格定义的概念去套用机器。“意识”“创造力”这些概念，其实人类自身也不理解。站在人类的角度上，去讨论 AI 有无意识这个问题，是自己陷入了一个思维的陷阱。因为究其根本，我们对自己有没有意识这件事也没有一个掌控，因而这样的讨论流于空泛。而对于这些概念的进一步掌握，取决于神经科学的进步。虽然不确认强人工智能是否很快会出现，但是即使出现，它也不一定需要以我们人类能理解的方式产生意识，达到目标。说不定在另一个外星观测者看来，我们也是无意识的，意识不过是这个被称作“人”的东西所使用的多级神经网络里某个调节参数的辅助工具。

AI 的运转方式和人脑相比一定是天壤之别的

这也是犯了太骄傲的毛病，因为你并不懂得人脑运算所采用的算法。人脑这个东西，即使是情感这些我们觉得很柔软的功能，背后也是以海量运算为背景的，而目前已被科学论文证实的是，在视皮层的运算，很多与目前的深度网洛运算很接近。有的人说人是向前看的动物，而机器只会向后看，事实上呢，人对未来的预测也来自于对过去数据的大量积累。

我们并无太多证据说明 AI 是否和我们的大脑有着相类似的运转方式，但有一点可以肯定的是，AI 的发展源自我们对自身的模仿，而对 AI 的探究反过来正在帮助我们理解我们自身，这也是生命最终的意义。

正如费曼所说，只有你知道如何制造一个东西时，你才真正理解它。

第五部分　人工智能应用谈

23 人工智能会取代艺术家的工作吗

当我们看到深度学习算法生成的绘画，谱写的音乐，甚至写的诗，我们意识到，也许有一天，算法将取代艺术家的工作。本文讲讲笔者对这个问题的思考，思考的起点是一句人尽皆知的话，“艺术来源于生活，但高于生活”。

先回答这样一个问题，对于计算机来说，哪一样更简单，模仿一副世界名画的风格，还是创造出一种全新的风格。给神经网络一副印象派的画，就可以生成印象派的变种，这不难实现，当前已有这样的应用。然而，如果只给对抽象派的文字描述，让算法生成一幅抽象派的绘画，目前还做不到，即使未来可以做到，但要做到开创一种全新的艺术流派，也必须要强人工智能才可行。因为无论哪种艺术，要想成家立派，都先必须观察生活的方方面面，看到自己当下身处时代的各式各样的人的生活，看到人群中普遍的焦虑和期待，然后以生活中逻辑上自洽的方式，超越性地创造出一种对当下困苦的暂时解脱。

无论是红楼梦中的大观园，还是第九交响曲中的欢乐颂，抑或是莫奈笔下的睡莲，都以自己独特的方式，超越了老天为创作者设下的极限，记录下了那个时代的真善美，从而使得作品和其背后的故事一起成为流芳百世的传说。强人工智能时代，艺术可以被看成、是一个受约束的优化问题，这不一定需要什么形而上的“意识”或者“灵魂”，但其需要的信息整合程度，却远远超越当前的艺术。

更关键的是，正是由于计算机不会感觉到痛苦，也不会有听力或者视觉上的障碍，所以不会像人类艺术家那样有一份沉重但甜蜜的枷锁。最好的艺术需要突破限制，而不是一开始就不受约束。计算机的强大运算能力，到了艺术领域，会因为无法产生一个像人类艺术家那样的故事而成为一种弱点。这也是未来艺术家的第一条路，就是创造带有个人故事的作品，让你的故事成为你的作品的唯一标识。

第二个问题，写一个专门生产抗日神剧剧本的程序容易，还是写一个能写出莎士比亚戏剧的程序容易。答案显然是前者，既然是神剧，那么台词、情节都有相应的套路，故事的剧情也没有多少意外。之所以要举这个例子，是因为当我们问什么时候艺术家会被算法取代时，要先问自己这个时代的审美品味是怎样的？高品位的艺术家更不容易被算法取代。更可能的一种情况是，算法通过将艺术中那些重复的机械的部分自动化，从而帮助艺术家更方便地创作，同时，更多的艺术家将时间和经历花在帮助普通大众提高艺术审美的品味上，让他们只喜欢那些只能由人类写出的歌曲或小说，而不是可以由机器代工的八股文。

音乐、文学诗歌这种抽象程度较高的艺术品种，有品味的高下之分，而有些艺术则不是。第三个问题，是我们最在乎艺术的是其在哪方面的价值，是拥有一件艺术珍品，还是创造出属于自己的艺术品，哪怕自己做得没那么好。在过去，好的艺术品是稀少的，而当人工智能使得艺术品不再成为稀缺品时，创造艺术的体验就变成真正有价值的东西。这指明了艺术家转型的出路，从自己创造艺术品，到让普通人去体验艺术创作的过程，画家可以让顾客去他的画室画一下午画，他教顾客去构图配色。厨师做饭也可以成为一种艺术创造的过程。当机器算法使得每个人都能拥有一幅过去的艺术名作时，艺术家们就可以教人们去画画、去学乐器。人的价值永远在于陪伴，在于创造体验。

总结一下，即使人工智能的兴起会给艺术家们带来一些挑战，但若是只站在存量上去思考，那么永远是新的威胁大于机遇的，只有站在增量的角度去看问题，我们才会意识到人工智能算法的兴起解放了艺术创作中的机械的成分，让艺术家可以帮助广大民众提高欣赏水平，让艺术家可以协助广大民众去体验艺术创作带来的心流体验。本文的思路也可以用在其他领域，面对人工智能的冲击，我们时该如何应对。

24 机器学习预测心理疾病

心理问题的诊断，一直都是难度很大的问题，我们培养专业的心理医生，但是专业的心理医生永远供不应求。大量得不到顾及的心理问题，不仅给个人和家庭带来巨大的损失，而且给企业和社会带来巨大的负担。

心理问题和各种身体问题一样，尽早发现永远比晚期发现好，但是由于我们往往很难发现自身的心理问题，而且也很少有人在问题不严重的时候有去看心理医生的习惯。

这时候 AI 可以帮助我们。一般你发现不了的东西，机器往往可以较早地发现。随着手机的普及和智能硬件的推广，我们获取各种人类行为和健康的数据越来越多，比如手环可以全面记录我们的心跳，手机可以记录我们的位置、我们的运动、我们的声音、我们社交活动的频率等。这些数据可以为我们提供各种心理问题的早期检测。

下面就介绍几个有趣的用机器学习来诊断心理疾病的方法。

第一个例子是心率测量。我们知道，最简单的生理测量是测心跳，通过心跳我们不仅可以知道身体健康状况，还可以测量心理压力。看似规律平稳的心跳，其实包含着众多信息，其中最重要的就是心跳变化率 HRV（Heart Rate Variability）。大家以为心跳是平稳周期性的，其实不然，如果仔细观察心跳每个时刻到每个时刻的变化，

会发现这里面在周期之外包含了丰富的变化。HRV 通常认为受人类非意识控制神经系统的影响，因而是一个认识神经系统的通道。我们在紧张、放松、工作、睡眠时 HRV 都会有变化。科学家设计了一组实验，实验人员让每个被测者（35 名 IT 从业者）带上一个胸带，睡觉时胸带可以记录被测者在睡觉时的状态，连续记录 4 个月时间，并回答标准问卷系统以确诊心理状态。实验人员发现了睡眠时 HRV 和工作时压力的相关性，提取了长短期的心跳变化率特征和频域特征，并用一个简单的 Logistic 模型测量了这个指数。最终测试可以达到 59%的准确率（低度、中度、高度压力分类任务，如果只分类高压和低压，则准确率可达到 66%；如果在更加可控的人群里，比如应考压力下的学生，则准确率可达 90%）。HRV 可以反映一个人长期承受压力的指数，虽然我们的大脑都有自愈能力，但是长期连续的压力会让这种自愈能力无法承担，从而造成多种心理疾病，如抑郁症。

第二个有趣的例子是社交网络测量抑郁症。为什么社交网络能测量抑郁症？首先，一个人在社交网络的行为往往反映其性格，这与神经症的发病率有直接的关系。

其次，一个人所表现的社交网络的行为变化往往体现他在一段时间的状态变化，自然语言识别可以比较有效地发现这些模式。2013 年的一篇论文讲述了作者通过 twitter 预测抑郁症的方法。这件事的历史最早可追溯到 2012 年 Park 发现人们有时会把自己接受抑郁症诊断和吃药的信息发到社交网络。之后 2013 年 Choudhury 分析了产前和产后妇女社交网络行为的变化，并根据产前的行为极好地预测了产后的行为，而这样的结果提示了我们这种方法的可行性。这篇文章的研究方法是抽取 476 个用户（一半男一半女，171 个有抑郁症的人和 305 个没有抑郁症的人）一年（对于患者是发病 1 年前的数据）的 twitter 数据并做特征工程。我们可以从 Engagement（参与度，各种发信息的频率和时间）、社交网络结构（中心性，双向性，自我趋向度）、情绪（正，负）和语言风格，以及抑郁高频词来诊断和预测抑郁症的发生。

一些基本的社交网络特征与抑郁症高度相关，图 24.1 显示了非抑郁症组（蓝色）和抑郁症组（红色）发状态的密度在一天内的变化，明显看出抑郁症的人更偏向于夜间活跃而正常组白天活跃。

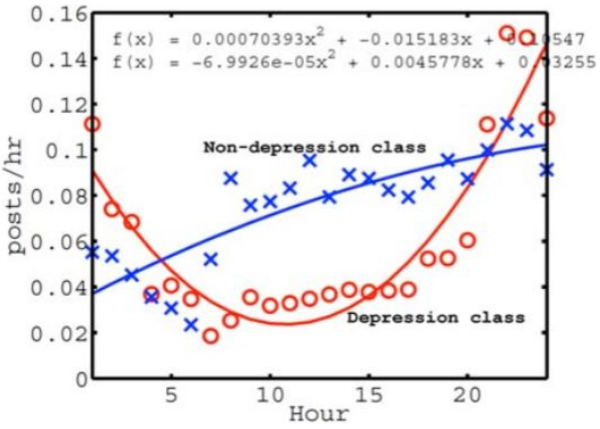


图 24.1

科学家利用这组数据总结了几种重要特征在一年时间里的变化，比如抑郁症患者社交网络的活动数量是递减的，但负面情绪表达和第一人称表达增加了，第三人称表达减少，以及抑郁高频词也增加了。

研究发现抑郁症患者的社交网络有封闭化趋势，趋向于自我中心，图 24.2 描写了用图论和网络中心度来测量抑郁的各项指数，包括向内、向外链接的个数、相互度、自我中心度等，最终可以看到抑郁症患者倾向于建立高度自我中心的小网络，在其中获取情感支持。

Egonetwork measures	Depres. class	Non-depres. class
#followers/inlinks	26.9 (σ=78.3)	45.32 (σ=90.74)
#followees/outlinks	19.2 (σ=52.4)	40.06 (σ=63.25)
Reciprocity	0.77 (σ=0.09)	1.364 (σ=0.186)
Prestige ratio	0.98 (σ=0.13)	0.613 (σ=0.277)
Graph density	0.01 (σ=0.03)	0.019 (σ=0.051)
Clustering coefficient	0.02 (σ=0.05)	0.011 (σ=0.072)
2-hop neighborhood	104 (σ=82.42)	198.4 (σ=110.3)
Embeddedness	0.38 (σ=0.14)	0.226 (σ=0.192)
#ego components	15.3 (σ=3.25)	7.851 (σ=6.294)

图 24.2

最终综合各项指标，社交网络分析可以预测下一段时间抑郁的发病率，正确率高达 70%，如图 24.3 所示。

	precision	recall	acc. (+ve)	acc. (mean)
engagement	0.542	0.439	53.212%	55.328%
ego-network	0.627	0.495	58.375%	61.246%
emotion	0.642	0.523	61.249%	64.325%
linguist. style	0.683	0.576	65.124%	68.415%
dep. language	0.655	0.592	66.256%	69.244%
demographics	0.452	0.406	47.914%	51.323%
all features	0.705	0.614	68.247%	71.209%
dim. reduced	**0.742**	**0.629**	**70.351%**	**72.384%**

图 24.3

第三个例子是利用 instagram 图片库来分析抑郁症的例子。科学家发现，正常人和患者在 instagam 上发布的照片有显著的色调区别，抑郁症患者明显对图像色调的偏好趋于清冷、灰白，而这种联系之强可以足够成为一种预测方法。

科学家可以通过抑郁和非抑郁被测人群喜好使用的图片滤镜判断其类型，如图 24.4 所示。

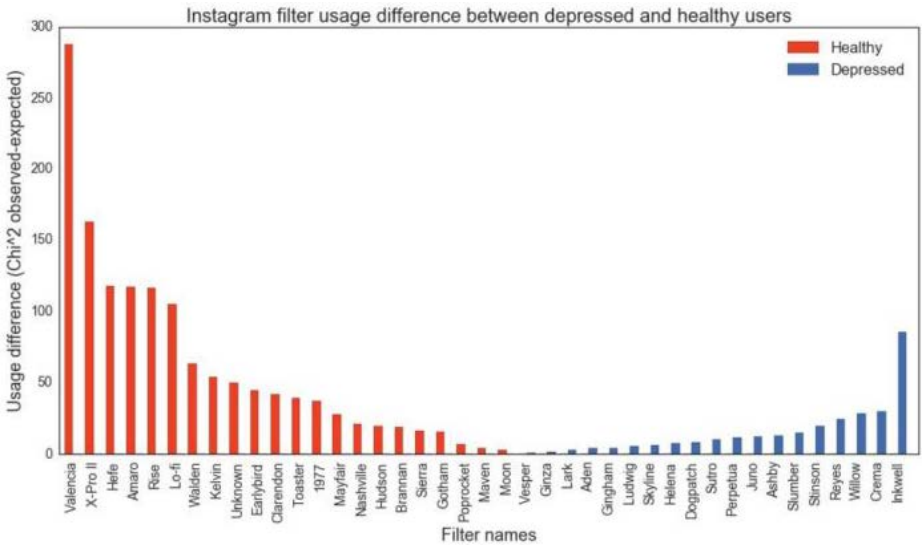


图 24.4

第四个例子是利用视线转移来分析自闭症。同样的图片，病患和正常人的视线转移是不同的。

第五个例子是声音检验，声音可以帮助我们检验多种心理疾病，以及心理压力。最早的通过声音检测压力的实验即使在室外手机记录下，这种检测精度也可以达到76%。声音检测方法几乎可以不通过理解语言本身的含义就知道你大脑的状态，所以也是一种比较成熟稳定的方法。

这些通过数据特征检验心理疾病的方法，为大数据辅助检测心理健康工作奠定了基础。

25 人机协作决策的两种方式

不管你是否愿意，机器都在帮你做决策。你在搜索引擎输入一个词语，会出现和该词有关的联想，你在 word 上码字，word 会给你划红线。这是机器在你做决策之后给你提的建议。然而，还有另外一种人机共同决策的方式，就是机器做出决策，然后人类用常识剔除那些明显有问题的决策。例如，Uber 的算法给出的建议是在大雪天涨价以激励司机多出来，但人类的审查者却可以判断这时候涨价对公司的形象不利，会激怒消费者，从而选择维持原价。

人类的决策，总是会受到种种认知偏差的影响，而这些偏差又是系统性的。将这些决策的数据教给机器，机器预测出你未来的可能决策也会带着偏见。这时算法就扮演了奸臣的角色，引诱你去选择更多会给你带来短期满足的东西。

如果反过来，人类在决策中扮演一个什么样的角色？答案是说服者。要知道决策永远不是一个人的事，也不是想明白了就结束的事情，让算法说服你该做什么决策，那就如同身为君王却成为了傀儡。没有人想成为机器的傀儡，所以即使未来机器做出的医疗诊断，会超过大部分的医生，但人们还是愿意和一个活生生的医生互动，因为只有人才能了解另一个同类的种种微妙的感受，才能说服别人。

对于未来的决策者，自己能否及时做出正确选择的能力已不再重要，真正重要的能力有三种：一是数据分析能力，即理解算法是如何做出决策的能力；二是体察他人

情绪等软性沟通技巧，三是深刻地理解人类的认知偏差，也就是人们常说的常识。什么叫深刻地理解，举个例子，你知道幸存者谬误的故事，这不算理解，你听到下面这个故事，也能想到幸存者谬误，才算是真正的理解了。一个出租车司机每天规定自己赚到一千元就回家休息，雨天他半天就赚到了，而晴天他需要跑 10 个小时。因而正确的做法是雨天多跑一会，晴天多休息一会，这样才能提高他每个工作时间内赚到的钱。

做到了上述三点的人，就是《未来简史》中说的超人。那些生活中的决策被 AI 决定了的人，即使给他们接上脑机接口，让他们拥有无限的信息，他们也一样无法说服别人，从而做出改变他人的决策。

26 小数据机器学习

巡洋舰科技的机器学习项目已展开一段时间，这几个月的一个突出感受是，人工智能的真正挑战是数据缺乏，我们面对的通常是小数据而非大数据问题。

人工智能如火如荼，很多公司想抓住人工智能的风口，然而在这些公司里，有很多人对数据的存留方式几乎可以用“落后”来形容，从几十万张纸片上的信息，到各路联络员脑子里的记忆……说实话，客户说他有十万组数据，清洗完能有一千行可用的就谢天谢地了。

比如说之前接的甲方预测农村土地报价的项目，甲方给我们做模型用的数据，大大小小清洗完，仅剩下六百多个可用的……这可用的里面还有一些用肉眼可去除的数据，比如北京某块十几亩的地一万块钱卖了，就算数据真实，对于我们想要构建的模型来说它也是噪音。

与大部分人想象的大数据时代不同，IT 时代走到现在这个阶段，大部分公司面临的是小数据问题，而非大数据。就算是携程这样种拥有巨量数据的公司，深入到某个具体问题时，依然面临的是小数据问题，比如云南的旅游线路推荐。那些预想的深度学习，比如人脸识别和 AlphaGo 中用到的卷积网络，是注定无缘这些问题的。那么这些领域是不是注定无法做机器学习呢？

数据小了，我们应该怎么办？其实小数据学习这件事，最会做的是我们人类自己。我们的人脑最擅长根据小量数据举一反三，通过形成对客观世界的常识和概念来学习。比如，俗语说的一朝被蛇咬，十年怕井绳，就是说被蛇咬了一次后（典型的小数据），然后就连看到井里的绳子，都会误以为那是蛇。掌握了规律，看到苹果落地，就懂得万有引力，所以能举一反三，而目前的大数据却不能。

所以目前一个征服小数据的有效手段，就是人为录入规则，把这种“人类的方法”加到机器学习里，从而大大减少需要学习的数据量。

好了，我们的解决方法是什么？直接做田野调查，在亲吻大地的同时，了解土地价格里面到底了藏了多少秘密。经调查发现，京郊普通土地 1 000 元亩/年比较正常，然而多出几倍叫价的情形也非常普遍，有些地方甚至十几倍叫价……这一观测使得我们团队果断选择了用“范围内准确率”替代原有的相对误差。

调查之后，与团队技术核心@子楠 商议，找出了一套更加适合小数据机器学习的方法。

第一种是贝叶斯方法，贝叶斯定律把已有的常识性信息通过先验概率嵌入公式里，而新的观测会修正先验得到后验概率，如果我们要通过为数不多的统计得到某城市的男女比例分布，那么把已知的全球男女比例 102:100 放入先验里，会大大加快少量数据下的结果稳定性。先验概率里包含的常识信息，可以大大减少我们需要的数据量。

实践中我们更多采用符合实践的旁门左道而非直接用贝叶斯公式，但是思想里无不包含着贝叶斯种子。

首先合理假设去除脏数据，越是小数据，个别的噪声对模型的影响就越明显。接着把土地转让价格转化成单价，然后根据在其他数据集上土地单价的相关知识，把价格大于 1600 元亩/年的都去掉。我们发现，在小数据问题上，像 Statistical Tests、Parametric Models、Bootstrapping 这类传统的方法依然十分有用。

接下来我们决定削减冗余特征，做足特征工程，最早网络抓取的特征 dummy 化（如性别变成男、女和中性三个特征，每个特征只取 0，1）后有 200 多个特征，如水井、农村土地补贴这些都纳入了因子特征，我们大胆地把特征合并后得到 10 个左右核心特征，如经济地理指数、交通指数、土地用途这一类，准确率比 200 多特征还高

一些。

当数据量较小时，复杂模型是灾难而不是福祉，不要过多特征，更不要特征之间的相互作用，有时神经网络跑出来的结果还不如线性回归。

这里我们用一种更激进的方法，就是数据生成法，又称模拟大法，就是在已有的原生数据里混入一些根据已有数据加上人为经验合成的人工数据，放在一起再来学习。通过生成数据，在原数据中强化了我们对客观规律的认知。

比如这块地要价 50 万元，一般会往高了要 10 万元，那么，预测 40~60 万元都认为是准确的。由于实际市场上人们偏向于报高价（万一瞎猫碰上死耗子呢）而非要低价，于是我们考虑人类本性，选择–33%到+50%的价格范围来做价格评估。这样准确率就轻松超越 40%，这反映了虽然存在大量瞎报价，但是合理报价依然还是占有很高比例的。

在调查中我们发现，人们报价的方法往往是按照单位面积的价格乘以面积（当土地上有房产时这种关系一定不成立，所以在数据统计中反而很难看出趋势），而价格随着年份呈现一定的贴现率关系。当年数增大一倍的，价格并不翻倍，而是以一定指数折损，这背后的原理很简单，我们对 10 年后的事情相对不关心，所以 10 年后的价值要用现在的钱来买，因而需要乘以一个折损率。

我们把这些客观经验加进去，以此来增加模型中的有效信息，进而提高模型效果。

首先提前拿出一部分数据藏起来作为测试集，然后再对训练集进行实验。比如我们的方法就是，复制一份训练数据出来，其他特征不变，面积全部增加或减少 1.1 倍，或年份增加或减少 1.1 倍，用这些生成的数据混合原生数据，然后看看准确率是否有提高。

我们发掘的规律从数据拟合中得到了印证，价格和年份的关系果真符合一定的贴现率，而这个比率是 5%，居然符合国家基本利率！果然经济规律无处不在。

最终我们所采集的各个数据集上的准确率都有提高，从 2%～10%，平均提高了 5%左右。

到目前为止，在甲方的 A 数据集上，利用对训练集加两倍数据的辅助方式进行

学习，每次重新随机测试，能够稳定达到 50%的准确率了。然而这种生成数据的思想，却是可以推衍的。

大量的中小型企业都存在这种数据少、质量差的情况，然而各个中小型企业之所以能生存，主要是因为大家脑海中有个“经验公式”。那么，把这个经验公式取出来，用某种方法来强化数据，这也许也是一条独特的人工智能化道路，毕竟我们的数据包含了太多特殊性。

大公司都在搞大数据，但是大多数公司还是只有小数据，如何利用好自己的经验，强化数据，运用机器学习，是这些小数据公司应重点思考的方向。

27 用深度学习玩图像的七重关卡

第一个重境界：图像识别

如果你正在了解深度学习的图像处理，那么你接触的第一个任务一定是图像识别：比如把你的爱猫输入一个普通的 CNN 网络里，看看它是猫咪还是狗狗。

一个最普通的 CNN，比如 CNN 的鼻祖 Lenet，如果你有不错的数据集（比如 kaggle 猫狗大战），就可以给出一个还差强人意的分类结果（80%左右的准确率）。

当然，如果再加上对特定问题的一些知识，也可以顺便识别一下人脸。

还可以识别一下猪脸，这样搞清楚每头猪的身份，对于高质量猪肉的销售，也是大有裨益的。

或者看看植物都有什么病害，不同的病斑它都可以看出来，这样做植物保护的人就可以只拿着手机下田了。

我们进化的方向，就是用更高级的网络结构取得更高的准确率。分类问题之所以简单，一要归功于大量标记的图像，二是因为分类的边界非常分明，即使机器不知道

什么是猫，什么是狗，但看出点区别还是挺容易的。如果你给机器几千几万类区分，那么机器的分辨力就下降了（再复杂的网络，在 ImageNet 那样分 1000 个类的问题里，都很难超过 80%的准确率）。

第二重境界：物体检测

很快你会发现，分类的技能在大部分的现实生活中并没有什么用。因为现实中往往是很多东西在一起，你拿猫狗大头照训练出来的分类网络一下子就乱了阵脚。即使是一张图片里只有一只猫和一只狗，给猫加点噪声，都可以使你的分类网络方寸大乱。

现实中，哪有那么多图片，一个图里就是一只猫或者一只狗的大图，更多的时候，一张图片里的东西往往又多又乱，没有什么章法可言，你需要自己做一个框，把你需要看的目标给框出来，然后，看看这些东西是什么。

于是你来到机器视觉的下一层挑战：目标检测（从大图中框出目标物体并识别）。随之而来的是一个新的网络架构，又被称为 R - CNN，即图片检测网络。这个网络不仅可以告诉你分类，还可以告诉你目标物体的坐标，即使图片里有很多目标物体，也可以一一找出来。

在众多路人甲中识别嫌疑犯，也是轻而易举，安防的人听着要按捺不住了。

2017 年出现的 YOLO 算法更是实现了快速实时的物体检测，你一路走过就告诉你视线里都有什么在哪里，这在无人驾驶里是何等的利器。

当然，到这里你依然会觉得无聊，即使网络已经很复杂，也不过是一个 CNN 网络（推荐区域），再加上一层 CNN 网络做分类和回归。能不能做点别的？

第三重境界：图像切割

啊哈，这就来到了第三个关卡，你不仅需要把图片中边边角角的物体给检测出来，还要把它们从图片中抠出来。要知道，刚出生的婴儿分不清物体的边界，比如桌子上有苹果这种事，什么是桌子，什么是苹果？所以，网络能不能把物体从一个图里抠出来，事关它是否真的像人一样掌握了视觉的本质。这也算是对它的某种“图灵测试”。把这个问题简化后，我们无非是在原图片上生成一个原图的“mask”，即面具，有

点类似于 Photoshop 里的蒙版。

注意，这个任务里，我们是要从一个图片里得到另一个图片，生成的面具是另一个图片，这时候，所谓的 U 型网络粉墨登场。注意，这是我们第一个生成式的模型。它的组成单元依然是卷积，但是却加入了 maxpooling 的反过程升维采样。

这个 Segmentation 任务的作用不可小瞧，比如现在私人卫星和无人机普及了，要不要去看看自己小区周围的地貌？轻轻输入，卫星图片一览无余。哪里有树，哪里有水，不需要人，全都可以抠出来。

即便你想数个细胞，也是挺容易的，给它变成这样的轮廓就可以了。

第四重境界：聚类

如果你是淘宝服装店的老板，想让客户输入一张服装的图片，然后得到一组推荐的服装，来个以图搜图的功能应怎么操作呢？注意，我可以从网络上爬一大堆图出来，但是这些数据是没有标注的，怎么办？

这里教你最简单的一招——聚类，也就是说，把图片统统放进卷积网络，但是我们不提取分类，而是只提取一些网络中间层的特征，这些特征有点像每个图片的视觉二维码，然后我们对这些二维码做一个 K-Means 聚类，也会得到意想不到的效果。为什么要深度？因为深度提取的特征，是与众不同的。

然后以图搜图呢？不过是找到同一聚类里的其他图片，在聚类的基础上做搜索。

第五层境界：降噪

我们开始晋升为仰望星空的人。作为一群仰望星空后观察细胞的人，我们最常发现的是我们得到的天文或者细胞图片的噪声实在是太大了，然后，深度学习给了你一套降噪和恢复图像的方法，一个叫 auto-encoder 的工具，起到了很大的作用，唰的一下，图像就清楚了。

这还不是最酷炫的，那个应用了博弈理论的对抗学习，也可以帮你去掉噪点！通过 GAN（一种图像生成工具），生成去掉噪点的图片，与没有噪点的自然图片，放在一起后，连卷积网络都判别不出来。

第六重境界：图像翻译

我们开始思考艺术、哲学，第一招，图像风格迁移。

真正能玩好这件事的，还是那个刚刚提过的对抗学习 GAN，比如大名鼎鼎的 CycleGAN，几乎可以实现一种你自定义的“图像翻译”功能，而且无须做标注，拿出冬天和夏天的两组图片，它会自动在两组图片中找出对应的。

第七重境界：生成图片

图像翻译之后，看是否能无中生有，在噪声里生成图片？

对，依然是 GAN，而且是最基础的卷积 GAN（DCGAN）就可以。

由此可见，GAN 是非常有前途的。

这里展示的七级浮屠，不过是深度学习被人类发现的冰山一角。

28 深度学习助力基因科技

最有可能改变未来的一项技术之一是基因科技，据麦卡锡 2016 年发布的一份报告，预计到 2025 年，全球将会累计产生 10 亿人次的全基因组数据，而检测基因的成本将下降到接近为 0。每一个人的全基因组数据将会达到 100GB，如何对这些海量数据进行解读，现有的生物信息方法基于统计学模型，而 2016 年 12 月 20 日在 bioRvix（生物领域的论文预印本）上，来自谷歌子公司 Deep Mind 的科学家发表了一篇论文，提出一个名为 DeepVariant 的检测工具，使用深度神经学习中的卷积神经网络 CNN 来检查基因组上的单碱基突变（SNP）和小的插入缺失（Indel），其准确性超越了当前主流的生物信息学软件 GATK。2016 年 12 月 30 日，来自康奈尔大学的另一组研究者又独立发表了一篇用类似方法检测基因变异的软件。两篇论文的爆发式出现，彰显了深度学习在基因组数据挖掘领域的巨大潜力。

每个人的体内都包含有 3 亿对碱基，这些碱基的排列方式决定了我们生老病死的方方面面，找到基因组的差异之处，可以帮助我们了解我们的疾病易感性，也可以知道我们的个体特征。对基因数据的传统分析方法，基于贝叶斯统计和专家的经验，针对不同的检测物种，需要很多人工的调节和试错。并且针对不同的测序平台，也需要进行特异性的配置。而基于深度学习的方法，巧妙地利用了图像识别领域成熟的技术，不仅可以以通用的学习流程达到现行软件相同的水平，更获得了由美国 FDA 监管的

SNP 检测算法比赛的第一名。这两项研究都还处于初期，考虑到深度学习算法巨大的提升空间，在可以预见的未来，基于深度学习的基因检测算法将会有更大的施展空间。

而在表观遗传学上，深度学习工具 DeepCpG 则可以根据基因序列预测单细胞层面是否会出现甲基化。在小白鼠的胚胎干细胞上验证后，DeepCpG 的表现优于现有的其他软件。甲基化会影响基因是否表达，而诸多启动子和增强子（DNA 序列中的调控区）与其调控的基因之间的相互作用则会影响基因的表达数量。同样是基于深度学习的 SPEID，基于基因序列预测启动子和增强子之间的相互作用，这是第一个基于基因序列来预测启动子和增强子的软件，其效果优于基于基因功能做出的预测。

传统的基因组关联分析（GWAS）大多只能检测一个点突变（SNP）与所研究疾病的关系，而 DeepWAS 这一新提出的框架则能够根据功能单元，选择出一组 SNP 的集合，来更加综合地研究治病的基因突变，并能直接地寻找调控区域的基因突变。在一项针对抑郁症的研究中，使用 DeepWAS 框架的新研究发现了一个新的控制抑郁症的主要基因：MEF2C。

以上列出的几项研究，都来自于 2016 年下半年，而在 2017 年 1 月，又有一篇基于 CNN 的文章尝试预测 HLA 基因与多肽的相互反应。这些文章虽然都来自于论文预印本，没有经过正规的同行审查，但不同领域的大规模井喷式爆发，也显示了该领域的巨大潜力。随着基因组数据的大量积累，以及深度学习开源平台的普及，未来深度学习必将成为生物信息领域的“一股清流”，为人类的健康做出贡献。

29 机器学习对战复杂系统

本书的最后，我们总结一下如何用机器学习应对复杂系统问题。

首先回顾一下复杂系统难以预测的原因。

（1）高维诅咒：构成现实生活的系统往往由大量未知变量决定。比如，生物由无数的细胞组成，基因由无数独立的单元组成，市场由无数的交易者组成，这些用物理的描述方法来预测，就是极高维度空间的运动问题。维度，首先使得再简单的方程形式都十分复杂难解。

此处补充维度的科学定义：维度是一个系统里可以独立变化的变量个数，一个有非常多变量的系统（如复杂网络），假如每个变量不是互相独立的，则也可以是低维系统。比如一个军营里的方阵，即使人数众多，但因为大家都做着一模一样的动作，所以只有一个独立变量，是一维系统。

（2）非线性诅咒。高维度系统的维度之间具有复杂的相互作用，导致我们不能把系统分解为单一维度，然后做加法的方法研究。高维加上非线性将得到对初级极为敏感的混沌系统。

非线性的一个重要推论是组织的产生，因为非线性，所以 1+1 既可以大于 2，也可以小于 2，为组织的产生提供了理论基础。

（3）反馈诅咒。复杂系统中反馈无处不在，即使是一个简单的一维系统，反馈也可以使得系统的特性很丰富，最典型的反馈是某种记忆效应，使得系统产生复杂的路径依赖，此刻你的现实与历史深刻关联，而关联方法又会产生复杂的模式。

反身性是一种由预测产生的特殊反馈，当你预测股市的价格时，会引起你的交易策略变化，从而影响你的预测，是为反身性。

（4）随机诅咒。复杂系统往往含有不包含确定规律的随机噪声，这些噪声使得系统的行为更加难以预测。很多时候，我们无法区分一个系统里发现的模式是由噪声导致的，还是由于元件之间的相互作用导致的。

这四大诅咒是这些系统难以理解和预测的原因，而这个时候，复杂系统和机器学习的方法论可以作为一种非常有力的手段帮助我们从复杂性中挖掘模式。

第一种方法叫模型驱动（Model Approch），即想办法找到事物变化的原因，用一种降维的思路列出微分方程，从非常繁复的要素中化简出最重要的一个或者两个，从而化繁为简，先抓住主要矛盾，其中的范例便是非线性动力学。

> 注：此处我们有两个基本假设让非线性动力学得到简化，一个是只讨论连续变量，另一个是不考虑系统内的随机性（无噪声项）。

（1）如果一个系统可以化简到一维，那么你只需要研究其内部存在的反馈性并描述它即可。负反馈导致稳定定点产生，正反馈导致不稳定性。很多事物都可以抽象为一维系统，包括简单环境下的人口增长问题。

（2）如果一个系统可以化简到二维，那么你需要研究两个维度间的相互作用，最终可以互为负反馈而稳定下来，互为正反馈而爆发，或者产生此消彼长的周期轨道。

（3）如果一个系统是三维的，则混沌可能产生。混沌即对初值极为敏感的运动体系。一旦偏离既定轨道一点，就几乎无法回去。

（4）如果一个系统大于三维，那么你需要用一个复杂网络描述它的运动，这个时候就可以得到我们复杂系统的主角——collective phenomena & emergence。复杂网络的性质主要取决于单体间相互作用的方式，以及系统与外界交换能量的方法，这两者

又息息相关。

复杂网络的动力学往往混沌难以预测，对于高维混沌系统，第一个方法是给出对事物定性的描述，然后祭出我们的第二个方法：先不管数据背后错综复杂的动因，而是直接以数据驱动我们的预测。

这其中的哲学内涵就是贝叶斯分析框架：即先不预测，而是列出所有可能的结果，并根据以往知识和经验预测每种结果发生的可能性（先验概率），之后不停吸收新的观测数据，调整每种可能结果的概率大小（后验概率），将想得到的结果概率最大化（MAP），最终做出决策。

如果把贝叶斯分析的框架自动化，让电脑完成，就会得到机器学习的最基本框架。

要想用机器学习来解决一个问题，往往需要具备三个条件：

（1）系统中可能存在模式。

（2）这种模式不是一般解析手段可以猜测到的。

（3）数据可以获取。

如果这三点有一点不符合，则很难运用机器学习。

机器学习的一个核心任务即模式识别，也可以看出它和刚才讲的复杂系统提到的模式的关系。虽然复杂系统难以通过其成分分析对整体进行预测，然而由于复杂系统通常存在模式，因此我们可以用模式识别来对系统进行归类，并预测各种可能的未来结果。

刚才讲的几大诅咒、高维、非线性、复杂反馈和随机性也被称为机器学习需要面对的几大核心难题，由此得到一系列机器学习的核心算法。

机器学习在现实生活中被用于许多方面，最常见的如商务洞察（分类、聚类、推荐算法），智能语音语义服务（时间序列处理、循环网络），各种自动鉴别系统，如人脸识别、虹膜识别、癌症检测（深度卷积网络），AlphaGo，机器人控制（深度强化学习算法）。而从方法论分，又可以分成有监督学习、无监督学习和强化学习。

贝叶斯决策的基本思想

要想让机器做决策，一个基本的思路是从之前统计的数据中挖掘已有的模式（Pattern）入手，来掌握新的数据中蕴含的信息。这个 Pattern 在有监督学习的例子里，就是把某种数据结构和假设结论关联起来的过程，我们通常用条件概率描述。那么让机器做决策，就是不停地通过新数据来调整这个数据结构（特征）与假设结果对应的条件概率。通常我们要把预先对某领域的知识作为预设（Prior），它是一个假设结果在数据收集前的概率密度函数，然后通过收集数据我们得到调整后的假设结果的概率密度函数，被称为后验概率（Posterior），最终的目标是让机器得到的概率密度函数与真实情况最匹配，即 Maximum a Posterior（MAP），这是机器学习的最终目标。

朴素贝叶斯分类器到贝叶斯网络

分类是决策的基础。在商业中，要根据收集客户的消费特征将客户分类，从而精准营销。在金融中，要根据一些交易行为的基本特征给交易者做分类。从贝叶斯分析的基本思路出发，我们可以迅速得到几种分类器。

（1）朴素贝叶斯分类器。它是机器学习中一个特别质朴而深刻的模型：当你要根据多个特征而非一个特征对数据进行分类时，我们可以假设这些特征相互独立（或者你先假设相互独立），然后利用条件概率乘法法则得到每一个分类的概率，最后选择概率最大的那个作为机器的判定。朴素贝叶斯分类器的基本框架如图 29.1 所示，c 是类别，A 是特征。

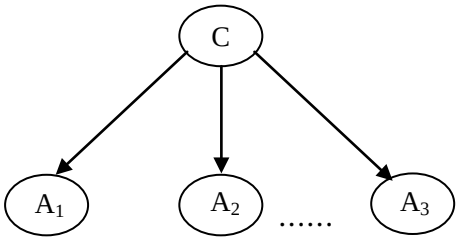


图 29.1

如果分类后的特征不是互相独立，而是互相之间有复杂的关联（这也是大部分时候我们所面临的问题），这时我们就需要更复杂的工具，即贝叶斯网络。比如对某些

病例的判定，咳嗽、发烧、喉咙肿痛都可以看作扁桃体发炎的症候，而这些症候又互为因果，此时贝叶斯网络是做出此类判定的最好方法。构建一个贝叶斯网络的关键是建立图模型，我们需要把所有特征间的因果联系用箭头连在一起，最后计算各个分类的概率。

贝叶斯网络对 MetaStatic Cancer 的诊断如图 29.2 所示，此处的特征具有复杂因果联系。

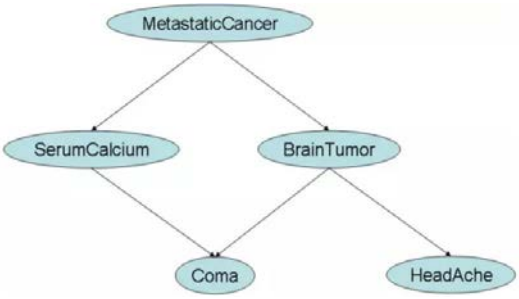


图 29.2

贝叶斯分析结合一些更强的假设，可以得到一些经常使用的通用分类器，比如 Logistic 回归模型，这里我们用到了物理里的熵最大假设，得到玻尔兹曼分布，因此之前简单贝叶斯的各个特征成立概率的乘积就可以转化为指数特征的加权平均。这是我们最常用的分类器之一。更加神奇的是，它在形式上同单层神经网络。

Logistic 函数如图 29.3 所示，数学形式同玻尔兹曼分布，是物理中熵最大模型的体现。

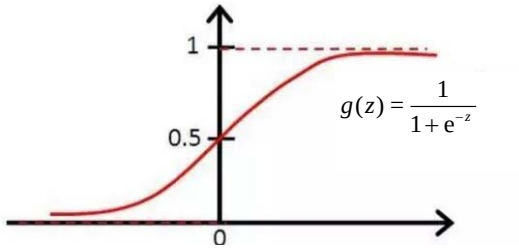


图 29.3

贝叶斯时间序列分析之隐马模型

贝叶斯时间序列分析被用于挖掘存储于时间中的模式，时间序列值是一组随时间变化的随机变量，比如玩牌的时候你的对手先后撒出的牌即构成一个时间序列。时间序列模式的预设（setting）即马尔科夫链。前面动力学模式里讲到，反馈导致复杂历史路径依赖，当这种依赖的最简单模式是下一刻可能出现的状态只与此刻的状态有关而与历史无关时，我们就能得到马尔科夫链。

马尔科夫链虽然是贝叶斯时间序列分析的基准模型，但是现实生活中遇到的时间序列问题，通常不能归于马尔科夫链，却可以间接地与马尔科夫链关联起来，这就是隐马过程，即含有隐变量的马尔科夫过程。

隐马过程示意图如图 29.4 所示。

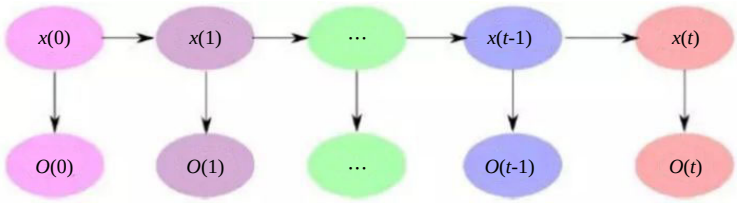


图 29.4

语音识别就是一类特别能利用隐马过程的应用，在这里语音可以看作一组可观测的时间序列，而背后的文字是与之关联的马尔科夫链。我们需要从可观测的量，按照一定的概率分布反推不可观测的量，并用马尔科夫链的观点对其建模，从而解决从语音到文字的反推过程。当今的语音识别则用到下面要讲的深度学习模型。

深度学习

刚刚讲的分类问题，只能根据我们已知的简单特征对事物进行分类，但假设我们手里的数据连需要提取的特征都不知道，那么应如何对事物进行分类呢？比如你要从照片识别人名，但你不知道选哪个特征和一个人关联起来。没关系，此时我们还有一个办法，就是让机器自发学习特征。通常在这类问题里，特征本身即构成一个复杂网络，下级的特征比较好确定，而最高层的特征，是由底层特征组合确定的，连我们人类自己都不能抽象出它们。

深度学习，即数据内涵的模式（特征）本身具备上述的多层级结构时，我们的机器学习机器也需要具有类似的多级结构，这就是大名鼎鼎的多层卷积神经网络。深度学习最大的优势是具有更高级的对“结构”进行自动挖掘的能力，比如它不需要我们给出所有的特征，而是自发去寻找最适合对数据集进行描述的特征。一个复杂模式，比如“人脸”，事实上可以看作一个简单模式的层级叠加，从人脸上的轮廓纹理这种底层模式，到眼睛、鼻子这样的中级模式，再到一个独特个体这种最高级的复杂模式，只有识别了底层模式，才有可能找到中级模式，而找到中级模式才方便找到高级模式，我们是不能从像素里一步到达这种复杂模式的，而是需要学习这种从简单模式到复杂模式的结构，多层网络的结构应运而生。

RNN 和神经图灵机

如果时间序列数据里的模式也包含复杂的多层级结构，这里和前面所说的复杂系统往往由于反馈导致复杂的时间依赖是一致的，那么要挖掘这种系统里的模式，通常的工具就是超级前卫的循环神经网络（RNN），这种工具对处理高维具有复杂反馈的系统有奇效，因为它本身就是一个高维具有复杂时间反馈的动力学系统。

循环神经网络，过去的信息可以通过循环存储在神经元之间，如图 29.5 所示。

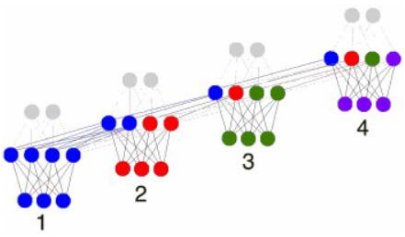


图 29.5

在一个复杂时间序列的问题里面，每个时间点的信息都可以对未来以任何方式产生复杂影响，这时处理这种复杂性的一个办法就是用循环神经网络，让它自发学习这种复杂结构。比如一个城市里的交通流，或者人与人之间的对话。

神经图灵机是在多层卷积神经网络或递归网络基础上加上一个较长期的记忆单元，从而达到处理需要更复杂时间关联的任务，比如对话机器人。而神经图灵机最厉害的地方在于，它可以通过机器学习传统的梯度下降法反向破译一个程序。比如你写

了一个 Python 程序，你用很多不同的输入得到很多对应的输出，你可以把它交给神经图灵机训练，最终本来对程序丝毫无所知的神经图灵机居然可以学会这个程序。

神经图灵机的基本结构如图 29.6 所示，在多级神经网络或循环网络（Controller）的基础上，增加一个更长时间的记忆，记忆和神经网络之间的读入和写出装置，整个装置都是可以被人工训练（可微分）的。

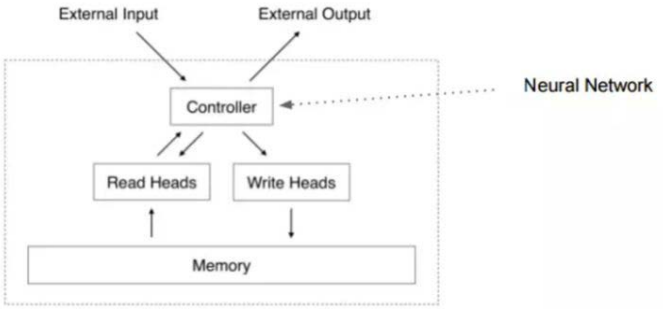


图 29.6

复杂科学解释模式的存在，机器学习解决模式的识别，两个学科加在一起就提供了我们面对复杂问题进行预测与决策的数学框架。世间的很多事情看似随机，却内藏规律，这种规律谓之模式。人类喜欢规律，喜欢在无规律中看出规律。一些真正隐藏在随机中的规律，很多时候会因为数据的维度太高，变化错综复杂而不可见。机器学习可以帮助我们看得清楚一点。如果我们只满足于模式识别，而不对模式进行解释（了解其动力学机制），则对于一些真正复杂的问题很容易流于表面，无法深入。比如我们通过对语言本身的统计和归类依然无法设计出和人相似的聊天机器人，而想让它最终和人相似，了解语言产生的动力学机制本身无疑是必要的。

注意，衡量风格之间的距离时，我们是把不同网络层级间的 gram 矩阵的距离都计算后加在一起，这样可以把不同层次度量的东西综合起来。

好了，到这一步，大功告成，把两个损失函数叠加在一起就可以了。

目标函数的设计学问很多，改变 a 和 b 的比例就能造成很多区别，大家注意，风格图片的比例越高，图像就越纹理化。

风格的权重变得很大时，图像就变成了意大利瓷砖！

然后我们可以做什么呢？梯度下降！注意，这里我们优化的目标不是网络权重，而是图像本身，这样就大功告成了！

当然这里说的只是风格迁移的一种，这种方法的优点是通俗易懂，缺点是速度很慢。还有一个方法，是借用生成网络，直接做出来，这个方法更快速，更加适合工业封装。

但是不要忘记，这件事给我们的启示绝不止这一个。它给我们的启示是，在我们深层神经编码的机制里，深度学习的踪影，你对风格的认知，其实和对内容的认知一样，是可以量化的，而不像某些艺术家所言，完全主观，与数学无关。不仅可以量化，而且这个信息是可以独立被提取的，这种信息不是存在于某个神经元之上，而是分布式地存在于多级神经网络的不同尺度之间，通过每一层神经元的统计规律表达。

虽然我们尚不知道这些猜想是否正确，但我们人类深奥的视觉处理机制提供了一种聪明的理解方法。